朱定局 著

数据处理与深度学习

清华大学出版社

数据处理与深度学习

朱定局　著

清华大学出版社
北　京

内 容 简 介

本书是一本原创性学术著作，主要内容包括深度学习模型与大数据智能处理的理论方法与应用实践。本书原创性在于提出、研究并给出了 14 种新型深度学习模型与大数据智能处理新理论、新方法和新应用。

本书可用于高等院校、科研院所及企事业单位的科研或作为教学用书。

图书在版编目(CIP)数据

数据处理与深度学习/朱定局著. —北京：清华大学出版社，2019
ISBN 978-7-302-51426-8

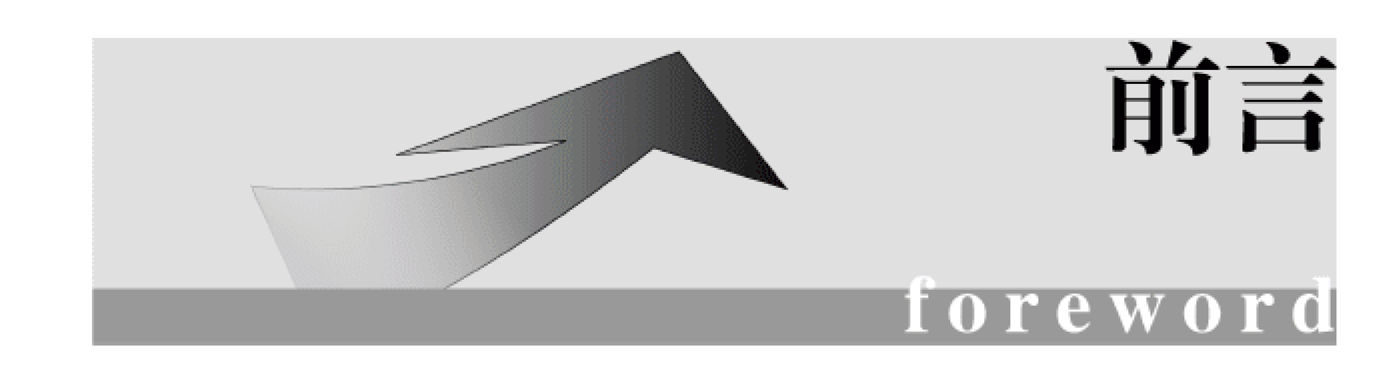
Ⅰ.①数…　Ⅱ.①朱…　Ⅲ.①数据处理—高等学校—教材　Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2018)第 242157 号

责任编辑：赵　凯　李　晔
封面设计：常雪影
责任校对：梁　毅
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：7.5　　**字　　数**：181 千字
版　　次：2019 年 6 月第 1 版　　**印　　次**：2019 年 6 月第1 次印刷
定　　价：49.00 元

产品编号：077768-01



前言

深度学习模型因为阿尔法狗战胜了人类围棋冠军而一举成名，再加上 BAT（百度、阿里巴巴、腾讯）、谷歌等国内外知名公司的推崇，使得深度学习模型的研究热浪一波高过一波。深度学习模型之所以能使得一度被打入冷宫的神经网络重新受到了大众的宠爱，主要因为深度学习模型能用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征，从而使得过去被废弃的大量非标签数据在神经网络中得到了利用；更为重要的是，使得神经网络的深度得到极大的增加，从而极大地提高了神经网络的处理能力。深度学习模型虽然是一个初生儿，但辈分极高，因为其前身是具有悠久历史的神经网络，在 20 年前，我上大学期间就有一门课程，称为“神经网络”。虽然辈分很高，但毕竟是初生儿，毕竟是一个新的生命，所以不论在理论上还是在应用上，都存在着很多盲点和缺陷，都需要开拓和创新，都需要完善和升级。

深度学习模型的春雨过后，无人驾驶、图像识别等人工智能相关的应用如雨后春笋层出不穷。随着应用的深度开展，其理论研究也在如火如荼地展开。很多企业和高校已经成立了深度学习研究中心或小组，专门研究深度学习模型的相关理论和应用，同时深度学习模型方面的培训和课程也在国内外掀起了热潮，但困境是深度学习模型理论和应用相关的著作仍然比较缺乏。

在深度学习模型火热之前，大数据已经热了几年，而深度学习模型的出现给大数据的智能处理提供了一个新的出路。深度学习模型只有基于大数据才能发挥其威力，才能达到很高的准确度。而大数据在应用了深度学习之后所能达到的处理效果也是传统算法所无法企及的。可以说，两者相得益彰，相互促进，互利共赢。

本书可以作为学习和研究深度学习模型的学生和学者的参考用书，其中的理论部分可以用于完善现有深度学习模型方法的不足，为理论的创新奠定基础；其中的应用部分可以为企业提供更多的深度学习模型应用思路和方案，进而使得深度学习模型在实践中产生更大的价值。

本书的所有内容都是作者原创性的研究成果。本书的原创性在于首次提出并研究给出了组合型深度学习模型、智能型深度学习模型、适应型深度学习模型；首次提出并研究给出了三种组合型深度学习模型，包括基于类别树的深度学习模型、基于联合聚类深度学习模型的数据识别、基于深度学习模型的身份识别；首次提出并研究给出了两种智能型深度学习模型，包括基于大数据的深度学习模型初始化、深度学习模型进化；首次提出并研究给出了两种适应型深度学习模型，包括期望与反期望深度学习模型、残缺深度学习模型；首次提出并研究给出了两种深度学习模型的预报预测大数据应用，包括基于大数据深度学习和分析

场的气象预报、分析预测文化冲突的深度学习模型；首次提出并研究给出了三种深度学习模型的调度推荐大数据应用，包括基于大数据和深度学习模型的云计算调度、基于大数据和深度学习模型的停车引导、基于大数据与深度学习的婚恋对象匹配推荐；首次提出并研究给出了两种深度学习模型的检测诊断大数据应用，包括基于深度学习模型和大数据的停车位检测、基于深度学习模型面向自动诊断的医疗数据处理。

本书选取的是人工智能的前沿领域和热点领域深度学习模型，但与各个应用领域进行了交叉创新，并在交叉创新的过程中发现和改进了现有深度学习模型理论的不足，同时拓展了现有深度学习模型的应用范围，进而实现了现有深度学习模型技术基础上从理论到应用的创新。

本书中的研究得到了国家级新工科研究与实践项目(粤教高函【2018】17 号)、国家社会科学基金重大项目(14ZDB101)、国家自然科学基金重点项目(41630635)、教育部—腾讯公司产学合作协同育人项目(201602001001)、广东高校重大科研项目(粤教科函【2018】64)、广东省新工科研究与实践项目(粤教高函【2017】118 号)、广东省高等教育教学研究和改革重点项目(粤教高函【2016】236 号)、广东省学位与研究生教育改革研究重点项目(粤教研函【2016】39 号)、广东省联合培养研究生示范基地(粤教研函【2016】39 号)的支持。

本书的特色在于以深度学习模型应用牵引深度学习模型理论的创新，同时以深度学习模型理论的创新促进深度学习模型应用的创新，在深度学习模型与各行各业之间进行了大胆的交叉创新，从而使得深度学习模型技术更智能、更普适、更精准、更强大，进而加速深度学习模型智能应用的发展。当然，深入后必能浅出，因此读者通过本书也能根据深度学习模型的新理论、新方法和新应用加深对现有深度学习技术的理解和应用。

华南师范大学
朱定局
2018 年 5 月

目录

contents



第1章

组合型深度学习模型

将现有技术中单个深度学习模型组合起来，能够优势互补，从而实现更为强大的功能或更好的性能。本章介绍了三部分内容：基于类别树的深度学习模型、基于联合聚类的深度学习模型的数据识别、基于深度学习模型的身份识别。基于类别树的深度学习模型，通过多级深度学习模型的逐层训练和逐层应用，降低了输入项数和输出类别数，从而降低了各级深度学习模型的结构复杂性及训练和应用难度，降低了深度学习模型的计算复杂度，同时提高了深度学习模型的准确率。基于联合聚类深度学习模型的数据识别方法，步骤如下：①获取 N 类数据样本集及对应的标签集并进行预处理，同时获取数据预设格式、标签预设格式；②对深度学习模型进行训练；③将每一类任一测试数据转化为该类数据预设格式后作为该类深度学习模型的输入，得到对应的测试输出标签；④根据测试输出标签所在标签集的元素个数及数据集之间的相似度计算，确定可能输出标签及最优输出标签；⑤计算各类输出标签一致和不一致的概率；⑥将可能、最优输出标签及上步中的概率输出。本技术通过相似度计算弥补了深度学习模型输出标签数量多而输入样本量不足时输出准确性低的不足，进而提高输出的准确性。基于深度学习模型的身份识别方法，步骤为：①获取 N 类输入数据；②初始化对应的 N 类深度学习模型；③训练 N 类深度学习模型；④根据采集成本从低到高对类别进行排序；⑤初始化 i、最优输出标签 L、最大相似度相对比值 U；⑥计算 T_i 类测试输出标签 L_i，并判断 L 是否与 L_i 相同；⑦获取 T_i 类相似度相对比值 U_i，并根据上一步的结果判别是身份识别失败还是执行下一步；⑧更新 L 和 U；⑨根据 U 与预设最大相似度相对比值 c 的关系、i 与 N 的关系，判别身份是否识别成功、是否需要将 i 加 1 继续循环执行。本技术优先选择低成本测试数据，并逐渐增加测试输入类型，在最低成本的情况下能够取得最高的判别准确率。

1.1 基于类别树的深度学习模型

深度学习模型通过输入数据和输出数据对深度学习模型进行训练[1,2]。传统的深度学习模型当输入项数和输出类别数过多时，就会导致输入数据及输出数据的量过大，从而使得深度学习模型的结构过于复杂；深度学习模型在训练时计算量过于庞大，甚至无法收敛，从而影响训练的质量，进而导致应用时的准确度降低。

综上所述，传统的深度学习模型当输入项数和输出类别数过多时准确度较低。

本节研究并给出了一种新的基于类别树的深度学习方法和神经网络系统，通过多级深度学习模型逐层训练和逐层应用，降低了输入项数和输出类别数，从而降低了各级深度学习模型的结构复杂性及训练和应用难度，降低了深度学习模型的计算复杂度，同时提高了深度学习模型的准确率。

1.1.1 基于类别树的深度学习方法

1. 基于类别树深度学习的基本方法

图 1-1 提供了一种基于类别树的深度学习方法，可包括以下步骤。

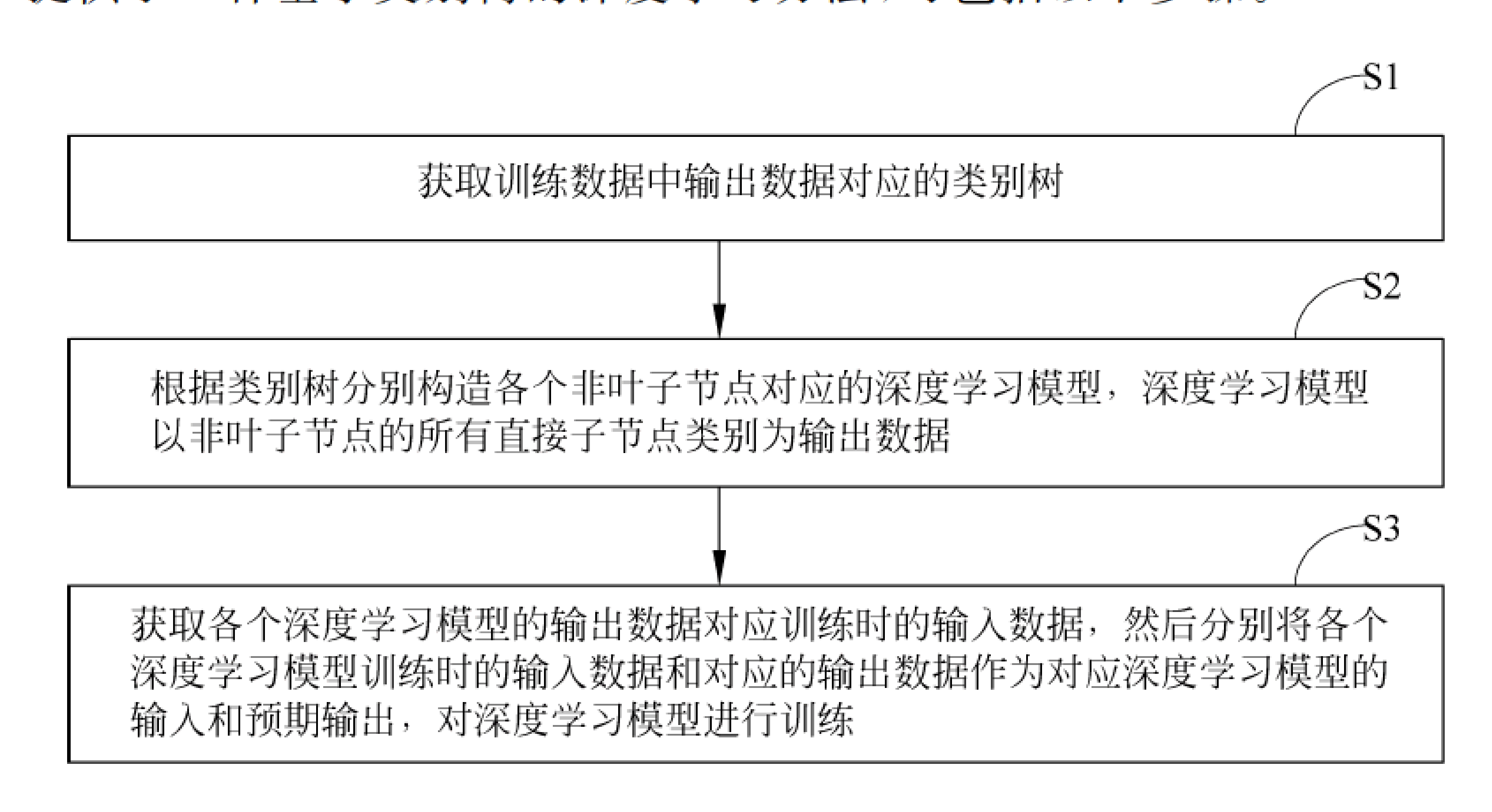


图 1-1 基于类别树的深度学习方法流程图

（1）步骤 S1，获取训练数据中输出数据对应的类别树。其中，获取类别树的方式可以是自动构建。

例如，一个技术方案的类别树如图 1-2 所示。输出数据包括但不限于输出标签。当训练数据中输出数据包括输出标签“猪”“狗”“草”“树”“藻”“男人”“女人”时，这些输出数据对应图 1-2 中类别树的叶子节点。例如，类别树中生物类别对应的训练数据中输出标签为“生物”；类别树中动物类别对应的训练数据中输出标签为“动物”。

在自动构建类别树时，可以对训练数据中的输出数据进行语义分析，获得对应的类别树。具体步骤为：①获取输出标签集合，将输出标签集合中每个输出标签作为一个叶子节

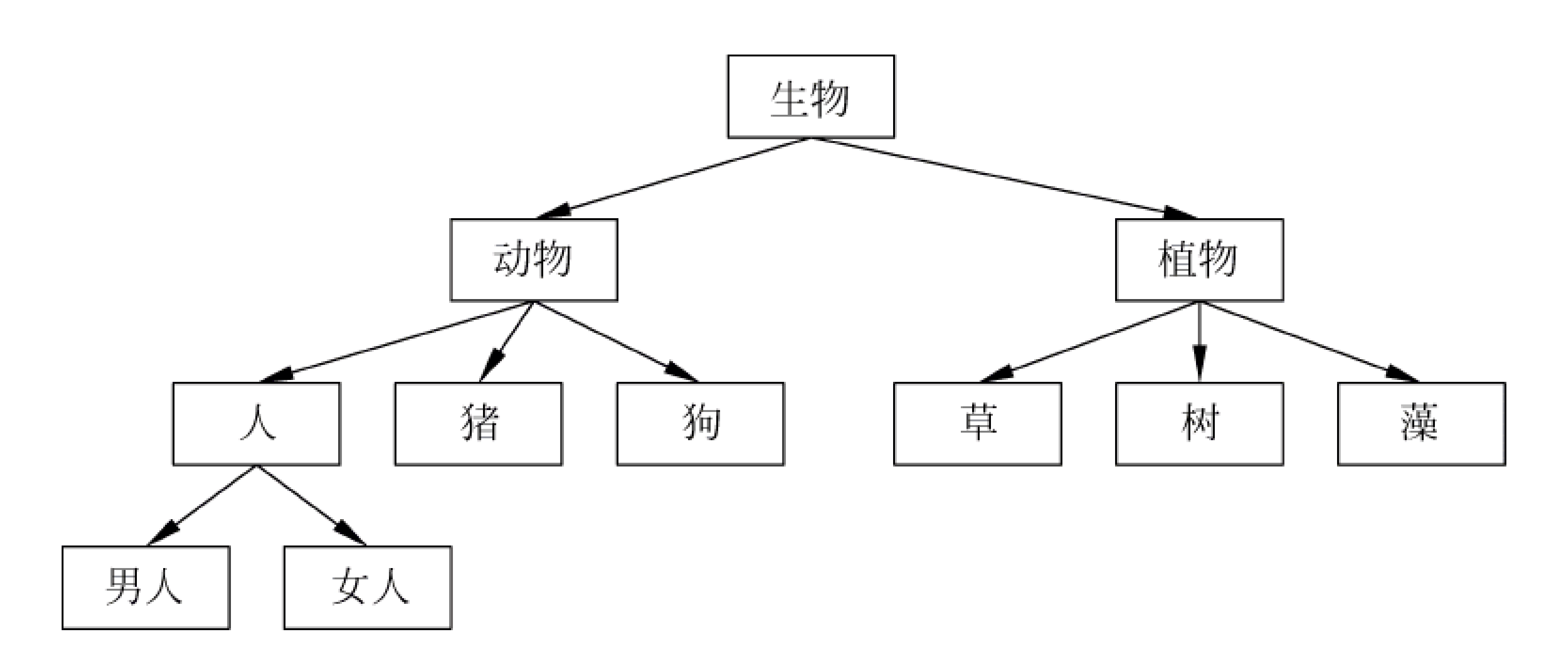


图 1-2　类别树的示意图

点；②从标签分类知识库中获取输出标签集合中每个输出标签所属的父类标签；③将属于同一个父类标签的各输出标签作为该父类标签的子类，并将该父类标签作为属于该父类标签的各输出标签的父节点；④将父类标签加入输出标签集合，将属于该父类标签的各输出标签从输出标签集合中删去，并返回步骤③，直至输出标签集合中只剩一个输出标签；⑤将剩余的输出标签作为类别树的根节点。

标签分类知识库是事先构建好的。

标签分类知识库中包括输出标签及其所属的父类标签。例如，输出标签“男人”“女人”属于父类标签“人”的子类；输出标签“猪”“狗”属于父类标签“动物”的子类；输出标签“草”“树”“藻”属于父类标签“植物”的子类。

标签分类知识库包括能区分父类标签的所有直接子类标签的特征变量集合。例如，能区分父类标签“植物”的所有直接子类标签“草”“树”“藻”类别的特征变量集合中包括特征变量“高度及颜色及纹理”；能区分父类标签“人”的所有直接子类标签“男人”“女人”类别的特征变量集合中包括特征变量“胸部”“头发”。

可以获取输出标签集合{猪，狗，草，树，藻，男人，女人}，将输出标签集合中每个输出标签作为一个叶子节点。

然后，从标签分类知识库中获取输出标签集合{猪，狗，草，树，藻，男人，女人}中每个输出标签所属的父类标签，“男人”“女人”属于“人”的子类；“猪”“狗”属于“动物”的子类；“草”“树”“藻”属于“植物”的子类。

将该父类标签“人”作为属于该父类标签的各输出标签“男人”“女人”的父节点，将该父类标签“人”加入输出标签集合，将属于该父类标签的各输出标签“男人”“女人”从输出标签集合中删去；将该父类标签“动物”作为属于该父类标签的各输出标签“猪”“狗”的父节点，将该父类标签“动物”加入输出标签集合，将属于该父类标签的各输出标签“猪”“狗”从输出标签集合中删去；将该父类标签“植物”作为属于该父类标签的各输出标签“草”“树”“藻”的父节点，将该父类标签“植物”加入输出标签集合，将属于该父类标签的各输出标签“草”“树”“藻”从输出标签集合中删去。得到新的输出标签集合{动物，植物，人}，然后重新从标签分类知识库中获取输出标签集合{动物，植物，人}中每个输出标签所属的父类标签，“动物”“人”属于“动物”的子类；“动物”“植物”属于“生物”的子类。

将该父类标签“动物”作为属于该父类标签的各输出标签“人”的父节点，将该父类标签“动物”加入输出标签集合，将属于该父类标签的各输出标签“人”从输出标签集合中删去，将

该父类标签“生物”作为属于该父类标签的各输出标签“动物”“植物”的父节点，将该父类标签“生物”加入输出标签集合，将属于该父类标签的各输出标签“动物”“植物”从输出标签集合中删去。得到新的输出标签集合{生物}，此时输出标签集合中只剩下一个输出标签，则将该输出标签{生物}作为类别树的根节点。

（2）步骤 S2，根据类别树分别构造各个非叶子节点对应的深度学习模型，深度学习模型以非叶子节点的所有直接子节点类别为输出数据。

具体步骤为：①以根节点为当前节点；②对以当前节点的所有直接子节点类别为输出数据时输出数据的格式为输出格式、以当前节点的所有直接子节点类别对应的输入数据的格式为输入格式对深度学习模型进行初始化，得到当前节点对应的深度学习模型；③分别将当前节点的各个直接子节点作为当前节点，并返回步骤②，直到当前节点为叶子节点为止。

其中，直接子节点指子节点自身，不包括子节点的子节点等更下一级子节点。例如，在图 1-2 中，节点“生物”的直接子节点为节点“动物”和节点“植物”，节点“动物”的直接子节点为节点“人”、节点“猪”和节点“狗”。

通过分级构造深度学习模型的方式，使得单个深度学习模型的输出数据的类别数量不会过多，从而降低了训练的复杂性，提高了应用时输出的准确度。

仍以图 1-2 所示的类别树为例，在本步骤中，可以先将“生物”节点为当前节点，当当前节点为“生物”节点时，“生物”节点的所有直接子节点包括“动物”“植物”。对以当前节点“生物”的所有直接子节点类别“动物”“植物”为输出数据时输出数据的格式为输出格式、以当前节点“生物”的所有直接子节点类别“动物”“植物”对应的输入数据的格式为输入格式对深度学习模型进行初始化，得到当前节点“生物”对应的深度学习模型。

（3）步骤 S3，获取各个深度学习模型的输出数据对应的训练的输入数据，然后分别将各个深度学习模型的训练的输入数据和对应的输出数据作为对应深度学习模型的输入和预期输出，对深度学习模型进行训练。

具体步骤为：①将以非叶子节点的每个直接子节点对应的类别作为对应深度学习模型的训练数据中的输出数据；②将非叶子节点的每个直接子节点类别对应的样本数据作为非叶子节点对应的深度学习模型的训练数据时的初始输入数据。通过分级使得级别越高（根节点的级别最低）的节点类别对应的训练数据越少，从而提高了训练的针对性，并能降低训练的复杂度。

例如，当前节点为“生物”节点时，“生物”节点的所有直接子节点包括“动物”“植物”。将以非叶子节点“生物”的每个直接子节点“动物”“植物”对应的类别作为对应深度学习模型的训练数据中的输出数据；将非叶子节点“生物”的每个直接子节点“动物”“植物”类别对应的样本数据作为非叶子节点“生物”对应的深度学习模型的训练数据的初始输入数据。此时训练时输入数据是动物图像，对应的预期输出是“动物”标签；训练时输入数据是植物图像，对应的预期输出则是“植物”标签。

或者在本步骤中，可以通过标签分类知识库获取能区分非叶子节点的各个直接子节点类别对应的各个直接子类标签的特征变量集合；非叶子节点对应父类标签；从初始输入数据中抽取特征变量集合对应的输入数据，作为对应深度学习模型的训练时的输入数据。

例如，当前节点为“生物”节点时，“生物”节点的所有直接子节点包括“动物”“植物”。从标签分类知识库中获取能区分当前节点“生物”的所有直接子节点“动物”“植物”类别的特征

变量集合为{颜色及纹理}，从初始输入数据动物图像和植物图像中抽取特征变量集合{颜色及纹理}对应的输入数据动物图像和植物图像，作为当前节点“生物”的深度学习模型的训练时的输入数据。

若标签分类知识库不包括能区分非叶子节点的各个直接子节点类别对应的各个直接子类标签的特征变量集合，则对属于非叶子节点的不同子节点类别的样本数据中各特征变量的数值进行比较；从各特征变量中选出在不同子节点类别对应的样本数据中数值差异大于预设阈值的特征变量集合，作为能区分当前节点的所有子节点类别的特征变量集合。本技术方案中的差异是指特征变量的数值的差别，例如，“草”头像的特征，包括颜色特征、纹理特征等。

例如，当前节点为“生物”节点时，“生物”节点的所有直接子节点包括“动物”“植物”。对属于当前节点“生物”的所有直接子节点“动物”“植物”类别的输入数据动物图像和植物图像进行比较，从中选出在不同子节点类别“动物”“植物”的样本数据动物图像和植物图像中各特征变量的数值进行比较，从各特征变量中选出在不同子节点类别“动物”“植物”对应的样本数据动物图像和植物图像中数值差异大于预设阈值的特征变量集合为{颜色及纹理}，作为能区分当前节点“生物”的所有子节点类别“动物”“植物”的特征变量集合。

在对当前节点对应的深度学习模型进行训练时，先利用训练数据中的输入数据对当前节点对应的深度学习模型进行无监督训练，再利用训练数据中的输入数据和对应的输出数据对当前节点对应的深度学习模型进行有监督训练。

例如，当前节点为“生物”节点时，将当前节点“生物”对应的深度学习模型的训练时的输入数据动物图像和植物图像以及输出数据“动物”标签、“植物”标签分别作为当前节点对应的深度学习模型的输入和预期输出，对当前节点对应的深度学习模型进行训练。或将当前节点“生物”对应的深度学习模型的训练时的输入数据动物图像颜色及纹理和植物图像颜色及纹理和输出数据“动物”标签、“植物”标签分别作为当前节点对应的深度学习模型的输入和预期输出，对当前节点“生物”对应的深度学习模型进行训练。

上述训练过程从根节点开始，以根节点为当前节点，对当前节点对应的深度学习模型进行训练，然后以根节点的各个直接子节点为当前节点，对当前节点对应的深度学习模型进行训练；重复上述过程，直到当前节点为叶子节点为止。

例如，当前节点为“生物”节点时，“生物”节点的所有直接子节点包括“动物”“植物”，分别以当前节点“生物”的每一个直接子节点“动物”“植物”作为当前节点重复上述训练过程，直至当前节点为叶子节点为止。

2. 基于类别树的深度学习的详细方法

在一个技术方案中，训练出深度学习模型之后，本技术方案包括以下步骤：①获取使用时的输入数据；②以根节点作为当前节点，以1作为当前级别，获取当前节点对应的深度学习模型；③获取使用时的输入数据对应于当前节点的深度学习模型的输出数据；④如果深度学习得到的输出数据的类别不是空标签，则继续执行以下两个步骤（如果深度学习得到的输出数据的类别是空标签，则不再执行以下两个步骤）：a. 将该类别作为使用时的输入数据在当前级别上所属的类别并输出，并且根据深度学习得到的输出数据的类别从类别树中选择该类别对应的节点；b. 以选择的节点作为当前节点，将当前级别加1，并返回步骤②，直到当前节点为叶子节点为止。

其中,获取使用时的输入数据对应于当前节点的深度学习模型的输出数据的步骤包括:将使用时的输入数据作为当前节点对应的深度学习模型的输入数据,通过深度学习得到输出数据;或者获取能区分当前节点的所有直接子节点类别的输入数据变量集合;将获取的输入数据变量集合对应的输入数据作为当前节点对应的深度学习模型的输入数据,通过深度学习得到输出数据。

使用深度学习得到的输出数据的类别作为使用时的输入数据在当前级别上所属的类别并输出,使用时能够获得多个级别的类别(对应于各级非叶子节点和叶子节点),相对于现有技术使用时只能输出一个级别的类别(对应于叶子节点)而言,输出结果更为丰富。例如,使用时输入数据为图像 A 在第 1 级别上所属的类别为“生物”,在第 1 级别上所属的类别为“生物”,在第 2 级别上所属的类别为“动物”,在第 3 级别上所属的类别为“人”,在第 4 级别上所属的类别为“男人”。

例如,可以获取使用时的输入数据为图像 A,初始时以根节点“生物”作为当前节点。当前节点为“生物”节点时,获取当前节点“生物”对应的深度学习模型。

当前节点为“生物”节点时,将输入数据图像 A 作为当前节点“生物”对应的深度学习模型的输入数据,通过深度学习得到的输出数据“动物”标签或“植物”标签或空标签,如果通过深度学习得到输出数据不是空标签,则将“动物”类别或“植物”类别作为使用时的输入数据图像 A 在当前级别上所属的类别并输出,并且根据深度学习得到的输出数据的“动物”类别或“植物”类别从类别树中选择该“动物”类别或“植物”类别对应的节点,以选择的“动物”类别或“植物”类别作为当前节点,将当前级别加 1。

或者,当前节点为“生物”节点时,获取能区分当前节点的所有直接子节点类别“动物”“植物”的输入数据变量集合{颜色及纹理}对应的输入数据图像 A 颜色及纹理作为当前节点“生物”对应的深度学习模型的输入数据,通过深度学习得到输出数据“动物”标签或“植物”标签或空标签,如果通过深度学习得到输出数据不是空标签,则将“动物”类别或“植物”类别作为使用时输入数据图像 A 在当前级别上所属的类别并输出,并且根据深度学习得到的输出数据的“动物”类别或“植物”类别从类别树中选择该“动物”类别或“植物”类别对应的节点,以选择的“动物”类别或“植物”类别作为当前节点,将当前级别加 1。

本技术方案根据深度学习得到的输出数据的类别来从类别树中选择该类别对应的节点,然后以该节点为当前节点重复上述深度学习过程,直至当前节点为叶子节点为止。

例如,当当前节点为“生物”节点,深度学习得到的输出数据的类别为“动物”标签时,以类别树中“动物”节点为当前节点重复上述深度学习过程,直至当前节点为叶子节点为止。

深度学习得到的输出数据的类别为“植物”标签时,以类别树中“植物”节点为当前节点重复上述深度学习过程,直至当前节点为叶子节点或深度学习得到的输出数据是空标签为止。

当当前节点为“动物”节点,深度学习得到的输出数据的类别为“人”标签时,以类别树中“人”节点为当前节点重复上述深度学习过程,直至当前节点为叶子节点为止。

深度学习得到的输出数据的类别为“猪”标签时,以类别树中“猪”节点为当前节点,当前节点“猪”为叶子节点,则停止重复上述深度学习过程。

深度学习得到的输出数据是空标签时,停止重复上述深度学习过程。

本技术方案还可以获取使用时输入数据对应的预期输出数据;如果预期输出数据与输

入数据经深度学习模型后得到的输出数据不一致，以使用时输入数据和预期输出数据作为训练数据，进一步地对各个深度学习模型进行训练。

例如，获取输入数据图像 A 在现实中对应的输出数据为“女人”，该输入数据图像 A 输入深度学习模型后通过深度学习得到的输出数据为“男人”，则以输入数据图像 A 及真实输出数据“女人”来作为训练数据对现有各深度学习模型采用上述同样方法进行训练。

上述基于类别树的深度学习方法，通过多级深度学习模型逐层训练和逐层应用，降低了输入项数和输出类别数，从而降低了各级深度学习模型的结构复杂性及训练和应用难度，降低了深度学习模型的计算复杂度，提高了深度学习模型的准确率。同时，使用深度学习得到的输出数据的类别作为使用时输入数据在当前级别上所属的类别并输出，能够获得多个级别的类别（对应于各级非叶子节点和叶子节点），相对于现有技术使用时只能输出一个级别的类别（对应于叶子节点）而言，输出结果更为丰富。

1.1.2　基于类别树的深度学习系统

如图 1-3 所示，本技术还提供一种深度学习系统，可包括：

（1）获取模块 1 用于获取训练数据中输出数据对应的类别树。

（2）构造模块 2 用于根据类别树分别构造各个非叶子节点对应的深度学习模型，深度学习模型以非叶子节点的所有直接子节点类别为输出数据。

（3）训练模块 3 用于获取各个深度学习模型的输出数据对应的训练时输入数据，然后分别将各个深度学习模型的训练时输入数据和对应的输出数据作为对应深度学习模型的输入和预期输出，对深度学习模型进行训练。

获取模块1 → 构造模块2 → 训练模块3

图 1-3　深度学习系统的结构示意图

上述深度学习系统，通过多级深度学习模型逐层训练和逐层应用，降低了输入项数和输出类别数，从而降低了各级深度学习模型的结构复杂性及训练和应用难度，降低了深度学习模型的计算复杂度，提高了深度学习模型的准确率。同时，使用深度学习得到的输出数据的类别作为使用时输入数据在当前级别上所属的类别并输出，能够获得多个级别的类别（对应于各级非叶子节点和叶子节点），相对于现有技术使用时只能输出一个级别的类别（对应于叶子节点）而言，输出结果更为丰富。

1.2　基于联合聚类深度学习模型的数据识别

现有深度学习模型能通过输入数据得到输出标签，例如，通过头像得到该人身份证号，或者通过语音得到该人身份证号。但在自顶向下的监督学习阶段必须要通过带标签数据的监督学习，例如，带有身份证号的头像，或者带有身份证号的语音。如果同时具备带有身份证号的头像和带有身份证号的语音，则可以将一个头像输入头像类对应的深度学习模型，从而得到输出的身份证号。将一个语音输入语音类对应的深度学习模型得到输出的身份证

号,然后判断通过输入头像和输入语音分别得到的输出的身份证号是否相同。如果相同则判断一个头像与一个语音对应的是同一人,如果不同则判断一个头像与一个语音对应的不是同一人。

但是,由于深度学习模型的准确率无法达到100%[3,4],所以一个头像输入头像类对应的深度学习模型后输出的身份证号可能是另一个头像相似的人的身份证号,一个语音输入语音类对应的深度学习模型后输出的身份证号可能是另一个语音相似的人的身份证号,那么就会导致不属于同一个人的一个头像和一个语音被判定为对应同一个人,也会导致属于不同人的一个头像和一个语音被判定为对应同一个人。当判断出一个头像和一个语音对应不同人时,却无法计算出该头像和该语音可能对应同一个人的概率;或者当判断出一个头像和一个语音对应同一个人时,却无法计算出该头像和该语音可能对应不同人的概率或其他人的概率。

当通过语音、头像,甚至更多其他类型的数据来识别对象时,使用现有深度学习技术,无法综合利用相似度及多种深度学习的结果来计算出其他可能性输出及最优输出,从而无法进行进一步的精准识别和判断。

本技术方案研究并给出了一种新的基于联合聚类深度学习模型的数据识别方法,步骤如下:①获取 N 类数据样本集及对应的标签集并进行预处理,同时获取数据预设格式、标签预设格式;②对深度学习模型进行训练;③将每一类任一测试数据转化为该类数据预设格式后作为该类深度学习模型的输入,得到对应的测试输出标签;④根据测试输出标签所在标签集的元素个数及数据集之间相似度的计算,确定可能输出标签及最优输出标签;⑤计算各类输出标签一致和不一致的概率;⑥将可能、最优输出标签及步骤⑤中的概率输出。

本技术方案通过相似度计算弥补深度学习模型在输出标签数量多而输入样本量不足时输出准确性的不足,进而提高输出的准确性。

1.2.1 基于联合聚类深度学习模型的数据识别方法

1. 基于联合聚类深度学习模型的 N 类数据识别

基于联合聚类深度学习模型的数据识别方法,包括以下步骤。

(1) 步骤 S1,先获取 N 类数据样本集及每一类数据样本集对应的标签集,再获取 N 类数据样本集中每一类数据样本的数据预设格式,并获取标签预设格式,然后对 N 类数据样本集和标签集进行预处理;$N\geqslant 1$。其中,获取 N 类中每一类数据样本的数据预设格式,并获取标签预设格式,具体步骤为:

获取每一类数据样本集中每一个数据样本的数据格式,将该类中相同的数据格式进行合并得到 s 种数据格式,统计该类数据样本集中每一种数据格式 P_i 对应的数据样本数量 M_i,将最大的 M_i 对应的数据格式 P_i 作为该类数据样本的数据预设格式;其中,$s\geqslant 1$,$i\geqslant 1$ 且 $i\leqslant s$。

获取每一类数据样本集对应的标签集中每一个标签的标签格式,将所有类中相同的标签格式进行合并得到至少 t 种标签格式,统计该类标签集中每一种标签格式 Q_j 对应的标签数量 N_j,将最大的 N_j 对应的标签格式 Q_j 作为标签预设格式;其中,$t\geqslant 1$,$j\geqslant 1$ 且 $j\leqslant t$。

对 N 类数据样本集和标签集进行预处理，具体步骤如下。

步骤 S1-1，判断每一类数据样本集中每一数据样本的数据格式是否与该类数据样本的数据预设格式一致，如果不一致，则将该类数据样本的数据格式转化为该类数据样本的数据预设格式；

步骤 S1-2，判断每一类数据样本集中每一个数据样本对应的标签的数据格式是否与标签预设格式一致，如果不一致，则将该类数据样本对应的标签的数据格式转化为标签预设格式；

步骤 S1-3，对 N 类数据样本集中的每类数据样本集进行聚类处理，获得 J 个聚类后的数据样本集及其对应的输出标签集；

步骤 S1-4，对 J 个聚类后的输出标签集中每个类中相同的标签进行合并，得到更新后的 J 个输出标签集；

步骤 S1-5，将更新后的 J 个输出标签集中具有相同标签的标签集及对应的数据样本集分别进行合并，得到预处理后的数据样本集及其对应的输出标签集。

（2）步骤 S2，初始化 N 类数据样本集对应的 N 个深度学习模型。具体为：

步骤 S2-1，将每一类数据样本的数据预设格式作为该类对应的深度学习模型的输入格式；

步骤 S2-2，将标签预设格式作为每一类对应的深度学习模型的输出格式；

步骤 S2-3，获取每一类对应的深度学习模型的配置信息，将其作为该类对应的深度学习模型的配置信息，然后对该类对应的深度学习模型进行配置。具体为：

① 从深度学习模型配置知识库中获取输入格式、输出格式与每一类数据预设格式及标签预设格式最为匹配的深度学习模型对应的配置信息，将其作为该类对应的深度学习模型的预设配置信息。

其中，

$$\begin{aligned}&\text{输入格式、输出格式与每类数据预设格式及标签预设格式的匹配度}\\&=\text{输入格式与该类数据预设格式的匹配度}\times u\%+\\&\quad\ \text{输出格式与标签预设格式的匹配度}\times(1-u\%)\end{aligned}$$

u 的默认值为 90。

② 将每一类对应的深度学习模型的预设配置信息输出给用户。

③ 获取用户对每一类对应的深度学习模型的预设配置信息的修改。

④ 将修改后的每一类对应的深度学习模型的预设配置信息，作为该类对应的深度学习模型的预设配置信息。

（3）步骤 S3，将步骤 S1 获取的每一类数据样本集作为输入，将其对应的标签集作为输出，对该类对应的深度学习模型进行训练，得到 N 个训练后的深度学习模型。具体为：

步骤 S3-1，将每一类数据样本集中每一个数据样本作为该类对应的深度学习模型的输入，对该类对应的深度学习模型进行自下而上的无监督训练。

步骤 S3-2，将每一类数据样本集中每一个数据样本作为该类对应的深度学习模型的输入，将该类数据样本集对应的标签集中该数据样本对应的标签作为输出，对该类对应的深度学习模型进行自顶向下的监督学习，得到 N 个训练后的深度学习模型。

（4）步骤 S4，为每一类对应的深度学习模型获取一个测试数据，将每一类测试数据的

数据格式转化为该类数据样本的数据预设格式，然后将该测试数据作为该类对应的深度学习模型的输入，通过该深度学习模型的计算得到该类对应的测试输出标签。

（5）步骤 S5，在步骤 S1 预处理的标签集中查找每一类的测试输出标签所在的标签集，之后判断该标签集是否只有一个标签元素。如果每一类的测试输出标签所在的标签集只有一个标签元素，则将每一类的测试输出标签作为该类的最优输出标签；否则执行下一步。

（6）步骤 S6，计算每一类的测试输出标签所对应的数据样本集与该类测试输出标签所在的标签集中每一标签元素对应的数据样本集的相似度，并根据该相似度计算并确定每一组可能输出标签；其中，每一组可能输出标签中包含了每一类的一个可能输出标签。具体为：

如果 $N=1$，则计算测试输出标签所对应的数据样本集与测试输出标签所在的标签集中每一标签元素对应的数据样本集的相似度，将相似度大于第一预设值 a 的所有标签元素作为一组可能输出标签；

如果 $N>1$，则获取第 i 类测试输出标签所对应的数据样本集 D_i；获取第 i 类测试输出标签所在的标签集中标签元素的个数 m_i；获取第 i 类测试输出标签所在的标签集中第 j 个标签元素对应的数据样本集 D_{ij}；计算 D_i 与 D_{ij} 的相似度 P_{ij}，其中，i 取从 $1\sim N$ 中的每一个自然数，j 取从 $1\sim m_i$ 中的每一个自然数。

对 $k_1, k_2, \cdots, k_N$ 的每一个取值，计算相似度第一综合值 $f(P_{1k1}, P_{2k2}, \cdots, P_{NkN})$，如果 $f(P_{1k1}, P_{2k2}, \cdots, P_{NkN})$大于第二预设值 b，则将第 1 类测试输出标签所在的标签集中第 k_1 个标签元素，第 2 类测试输出标签所在的标签集中第 k_2 个标签元素，……，第 N 类测试输出标签所在的标签集中第 k_N 个标签元素，作为一组可能输出标签；其中，k_1 取从 $1\sim m_1$ 中的每一个自然数，k_2 取从 $1\sim m_2$ 中的每一个自然数，……，k_N 取从 $1\sim m_N$ 中的每一个自然数；$f(P_{1k1}, P_{2k2}, \cdots, P_{NkN})$表示求$(P_{1k1}, P_{2k2}, \cdots, P_{NkN})$的乘积。

（7）步骤 S7，计算每一组可能输出标签中每一类的可能输出标签对应的数据样本集与该类测试数据集的相似度，并根据该相似度计算并确定一组可能输出标签作为最优输出标签。具体为：

如果 $N=1$，则计算每一组可能输出标签对应的数据样本集与该类测试数据集的相似度，获取最大相似度对应的一组可能输出标签作为最优输出标签；

如果 $N>1$，则计算每一组可能输出标签中第 i 类可能输出标签对应的数据样本集与该类测试数据集的相似度 P_i，之后计算相似度第二综合值 $g(P_1, P_2, \cdots, P_N)$，获取最大相似度第二综合值对应的一组可能输出标签作为最优输出标签；其中，$g(P_1, P_2, \cdots, P_N)$表示求$(P_1, P_2, \cdots, P_N)$的乘积，i 取 $1\sim N$ 中的每一个自然数。

（8）步骤 S8，计算可能输出标签中各类输出标签都一致及不一致的概率，作为各类输出标签一致及不一致的概率。具体为：

如果 $N=1$，则只有一类输出标签，因此各类输出标签一致的概率为 100%，不一致的概率为 0；

如果 $N>1$，则先判断每一组可能输出标签中各类可能输出标签是否一致；然后将判断结果为一致的各组可能输出标签对应的相似度第二综合值之和与所有可能输出标签对应的相似度第二综合值之和相除，得到各类输出标签一致的概率；最后以 100%减去各类输出标

签一致的概率得到各类输出标签不一致的概率。

（9）将可能输出标签、最优输出标签、各类输出标签一致及不一致的概率输出。

本技术方案与现有技术相比，其显著优点为：本技术方案将深度学习模型与相似度计算进行有机结合，丰富了输出的结果，提高了输出的准确率。本技术方案采用结合相似度计算的方法，取长补短，从而通过相似度计算来弥补深度学习模型在输出标签数量多而输入样本量不足时输出准确性的不足，进而提高输出的准确性。

以两类数据样本集及其对应的标签集为例。第一类数据样本集为{头像 11，头像 12，……，头像 1m}，对应的第一类输出标签的集合为{身份证号 11，身份证号 12，……，身份证号 1m}，其中，头像 11 与身份证号 11 对应，头像 12 与身份证号 12 对应，……，头像 1m 与身份证号 1m 对应。其中可能会存在相同的身份证号，例如身份证号 13 与身份证号 16 相同。第二类数据样本集为{语音 21，语音 22，……，语音 2n}，对应的第一类输出标签的集合为{身份证号 21，身份证号 22，……，身份证号 2n}，其中，语音 21 与身份证号 21 对应，语音 22 与身份证号 22 对应，……，语音 2n 与身份证号 2n 对应。其中可能会存在相同的身份证号，例如身份证号 22 与身份证号 28 相同。

2. 基于联合聚类深度学习模型的两类数据识别

下面以两类数据样本集为特例，结合图 1-4 进行说明。本技术方案基于联合聚类深度学习模型的数据识别方法，包括以下步骤。

（1）步骤 S1，首先获取两类数据样本集及每类数据样本集对应的标签集，再获取两类数据样本集中每一类数据样本的数据预设格式，并获取标签预设格式。具体为：

获取每一类数据样本集中每一个数据样本的数据格式，将该类中相同的数据格式进行合并得到 s 种数据格式，统计该类数据样本集中每一种数据格式 P_i 对应的数据样本数量 M_i，将最大的 M_i 对应的数据格式 P_i 作为该类数据样本的数据预设格式；其中，$s\geqslant 1$，$i\geqslant 1$ 且 $i\leqslant s$。例如，第 1 类数据样本为图像样本，第 2 类数据样本为语音样本。以第 1 类数据样本为例，第 1 类数据样本集中 480 像素×640 像素 JPEG 图片数据格式的数据样本有 809 个、480 像素×640 像素 TIFF 图片数据格式的数据样本有 8367 个、480 像素×640 像素 BMP 图片数据格式的数据样本有 67 个、2576 像素×1932 像素 JPEG 图片数据格式的数据样本有 5362 个、2576 像素×1932 像素 TIFF 图片数据格式的数据样本有 32 个、2576 像素×1932 像素 BMP 图片数据格式的数据样本有 136 个，其中数据样本数量最大的数据格式是 480 像素×640 像素 TIFF 图片数据格式，所以将 480 像素×640 像素 TIFF 图片数据格式作为第 1 类数据样本的数据预设格式。

获取每一类数据样本集对应的标签集中每一个标签的标签格式，将所有类中相同的标签格式进行合并得到至少 t 种标签格式，统计该类标签集中每一种标签格式 Q_j 对应的标签数量 N_j，将最大的 N_j 对应的标签格式 Q_j 作为标签预设格式；其中，$t\geqslant 1$，$j\geqslant 1$ 且 $j\leqslant t$。例如，共有两类数据样本集对应的标签集，第 1 类数据样本集对应的标签集中，身份证号标签有 5636 个，姓名标签有 5426 个；第 2 类数据样本集对应的标签集中，身份证号标签有 2654 个，姓名标签有 235 个，则两类数据样本集对应的标签集中身份证号标签有 8290 个，姓名标签有 5661 个，所以将身份证号标签作为标签预设格式。

然后对输入的两类数据样本集和标签集进行预处理，结合图 1-5，以第 1 类数据样本集和标签集为例，具体过程为：

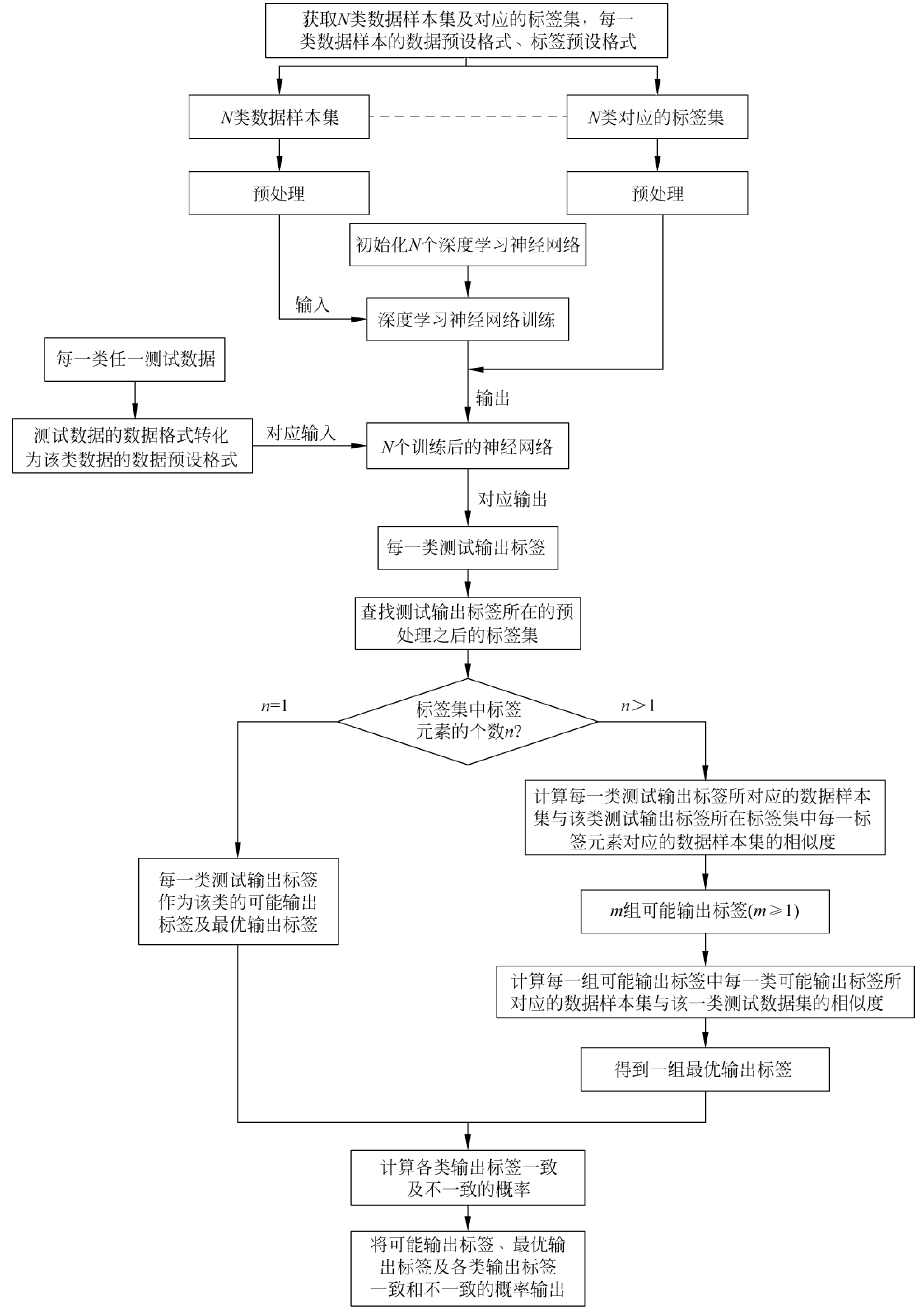


图 1-4　基于联合聚类深度学习模型的数据识别方法流程

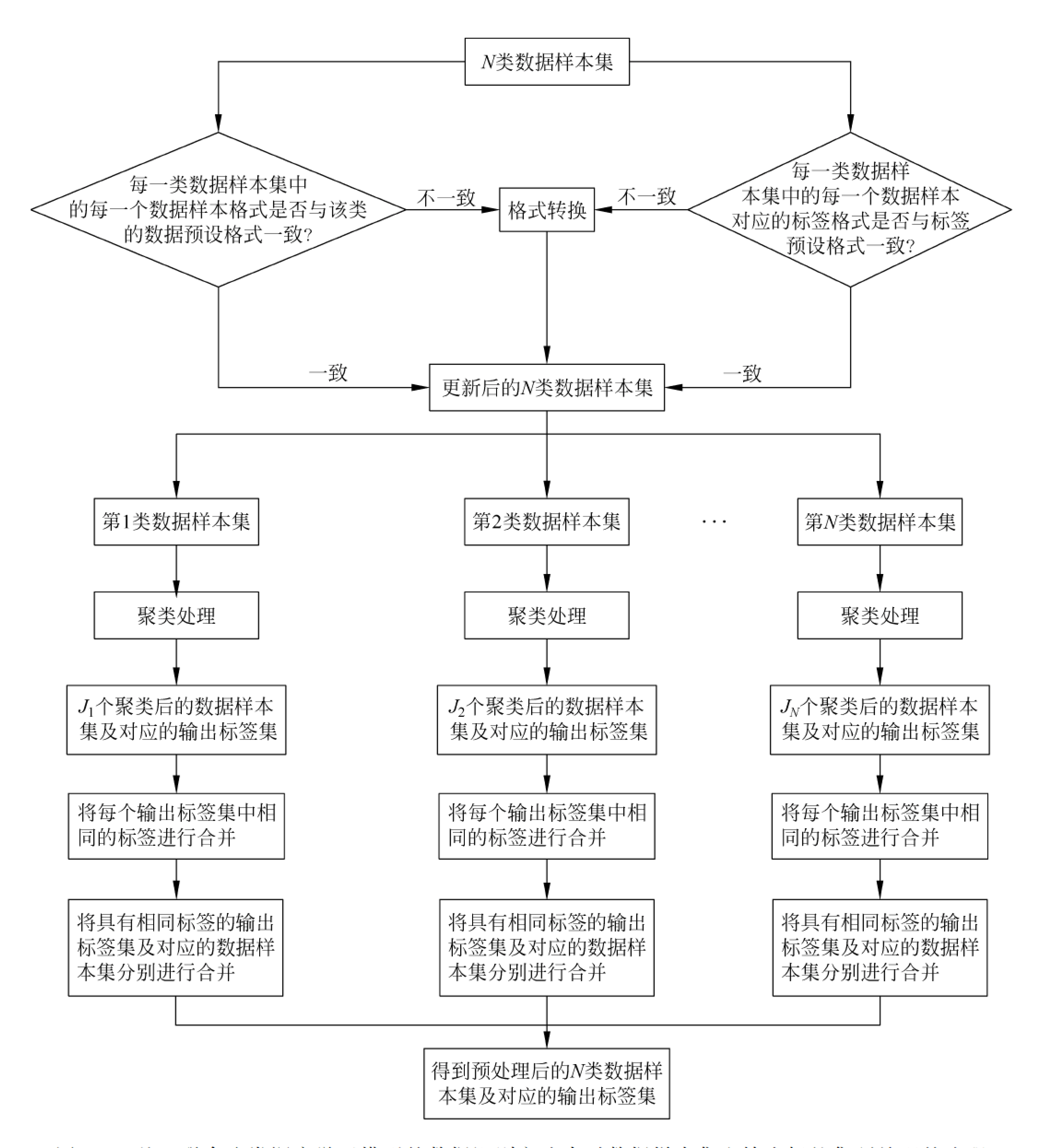


图 1-5 基于联合聚类深度学习模型的数据识别方法中对数据样本集和输出标签集预处理的流程

步骤 S1-1,判断每一类数据样本集中每一数据样本的数据格式是否与该类数据样本的数据预设格式一致,如果不一致,则将该类数据样本的数据格式转化为该类数据样本的数据预设格式。例如,480 像素×640 像素 TIFF 图片数据格式是第一类数据样本的数据预设格式,如果第一类数据样本集中一个数据样本的数据格式也是 480 像素×640 像素 TIFF 图片数据格式,与第一类数据样本的数据预设格式相同,则无须进行转化;如果第一类数据样本集中一个数据样本的数据格式是 2576 像素×1932 像素 JPEG 图片数据格式,与第一类数据样本的数据预设格式不同,则须转化为 480 像素×640 像素的 TIFF 图片数据格式。

步骤 S1-2,判断每一类数据样本集中每一个数据样本对应的标签的数据格式是否与标签预设格式一致,如果不一致,则将该类数据样本对应的标签的数据格式转化为标签预设格

式。例如，身份证号标签作为标签预设格式，如果第一类数据样本集中一个数据样本对应的标签的数据格式是身份证号格式，与标签预设格式相同，则无须进行转化；如果第一类数据样本集中一个数据样本对应的标签的数据格式是姓名格式，与标签预设格式不同，则须转化为身份证号格式。

步骤 S1-3，对第 1 类数据样本集进行聚类处理，获得 j 个聚类后的数据样本集及其对应的输出标签集。具体为：

首先对第 1 类数据样本集{头像 11，头像 12，……，头像 $1m$}进行聚类，聚类的规则是：将相似度大于头像相似度预设阈值（默认值为 90%）的头像加入到同一个聚类（即将上述头像内部进行相似度计算，如果相似度大于 90%，则将相应的头像加入一个聚类），使得一个聚类中任一头像都在该个聚类中存在另一头像与该个头像的相似度大于头像相似度预设阈值（默认值为 90%），同时使得一个聚类中任一头像都在另一个聚类中不存在任一头像与该头像的相似度大于头像相似度预设阈值（默认值为 90%）；每一头像只属于一个聚类。根据该聚类规则，得到第一个聚类{头像 111，头像 112，……，头像 $11m_1$}，第 2 个聚类{头像 211，头像 212，……，头像 $21m_2$}，……，第 j 个聚类{头像 j_{11}，头像 j_{12}，……，头像 j_{1m_j}}。

然后对第 1 类数据样本集对应的标签集进行聚类。第 1 类数据样本集的第 1 个聚类{头像 111，头像 112，……，头像 $11m_1$}对应的输出标签的集合为{身份证号 111，身份证号 112，……，身份证号 $11m_1$}，作为第 1 类输出标签的第 1 个聚类；数据样本集的第 2 个聚类{头像 211，头像 212，……，头像 $21m_2$}对应的输出标签的集合为{身份证号 211，身份证号 212，……，身份证号 $21m_2$}，作为第 1 类输出标签的第 2 个聚类；……；数据样本集的第 j 个聚类{头像 j_{11}，头像 j_{12}，……，头像 j_{1m_j}}对应的输出标签的集合为{身份证号 j_{11}，身份证号 j_{12}，……，身份证号 j_{1m_j}}，作为第 1 类输出标签的第 j 个聚类。

步骤 S1-4，对上述 j 个聚类后的输出标签集中每个类中相同的标签进行合并，得到更新后的 j 个输出标签集。

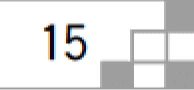
步骤 S1-5，将更新后的 j 个输出标签集中具有相同标签的标签集及对应的数据样本集分别进行合并，得到预处理后的数据样本集及其对应的输出标签集。

例如，第 1 类数据样本集的第 1 个聚类{头像 111，头像 112，……，头像 $11m_1$}对应的输出标签的集合为{身份证号 111，身份证号 112，……，身份证号 $11m_1$}；数据样本集的第 2 个聚类{头像 211，头像 212，……，头像 $21m_2$}对应的输出标签的集合为{身份证号 211，身份证号 212，……，身份证号 $21m_2$}；如果第 2 个聚类对应的输出标签的集合中身份证号 212 与第 1 个聚类对应的输出标签的集合中身份证号 116 相同，则将第 2 个聚类对应的输出标签的集合与第 1 个聚类对应的输出标签的集合合并，同时将第 2 个聚类对应的数据样本集与第 1 个聚类对应的数据样本集合并。

（2）步骤 S2，初始化 N 类数据样本集对应的 N 个深度学习模型。

（3）步骤 S3，将步骤 S1 获取的每类数据样本集作为输入，将其对应的标签集作为输出，对该类对应的深度学习模型进行训练，得到两个训练后的神经网络。结合图 1-6，具体为：

步骤 S3-1，将每类数据样本集中每一个数据样本作为该类对应的深度学习模型的输入，对该类对应的深度学习模型进行自下而上的无监督训练。



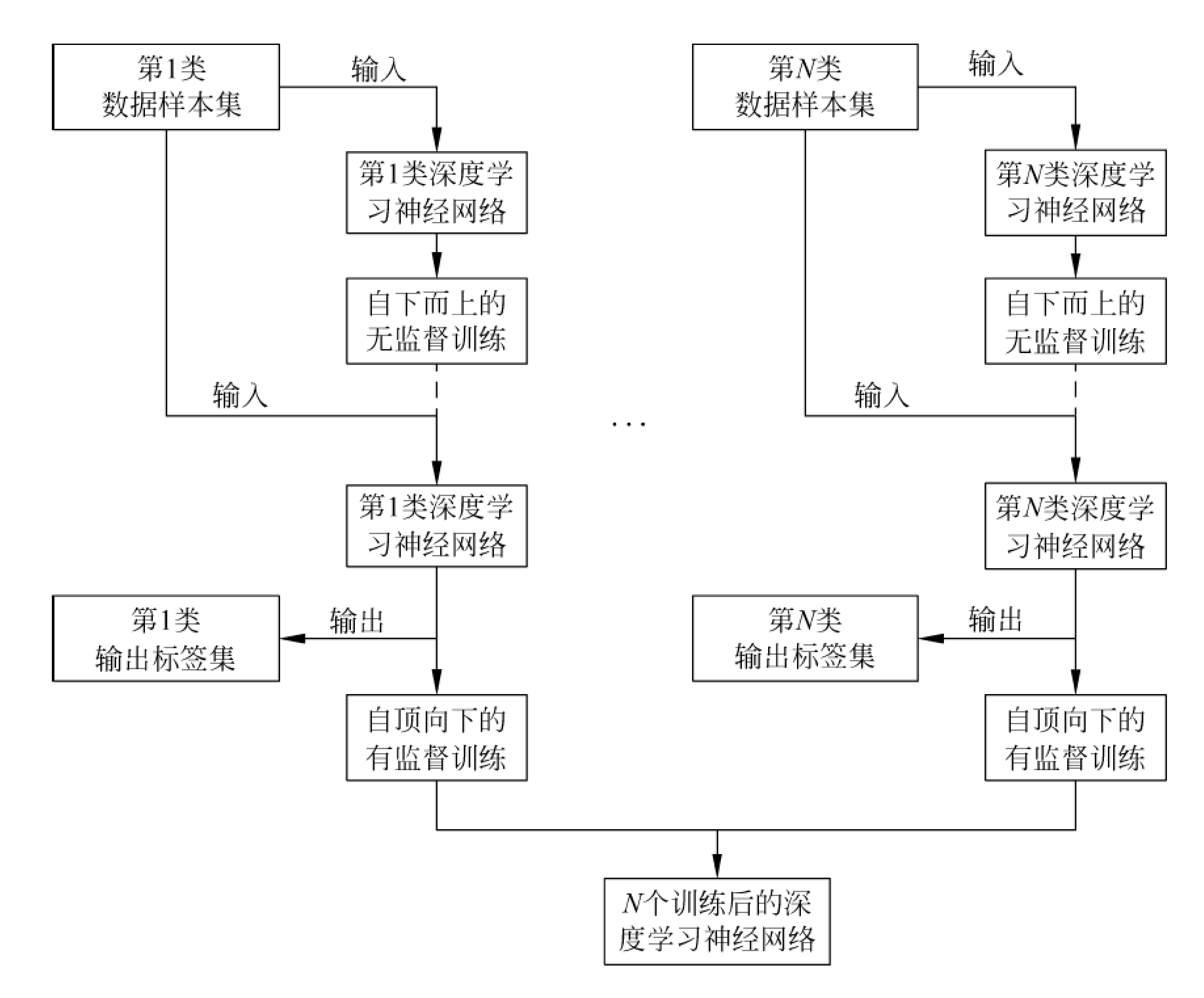


图 1-6 基于联合聚类深度学习模型的数据识别方法中深度学习模型训练的流程

步骤 S3-2,将每类数据样本集中每一个数据样本作为该类对应的深度学习模型的输入,将对应的标签集中的对应标签作为输出,对该类对应的深度学习模型进行自顶向下的监督学习,得到两个训练后的深度学习模型。

步骤 S3-3,为每一类对应的深度学习模型获取一个测试数据,将每一类测试数据的数据格式转化为该类数据的数据预设格式。例如,480 像素×640 像素 TIFF 图片数据格式是第 1 类数据样本的数据预设格式,如果第 1 类该测试数据的数据格式也是 480 像素×640 像素 TIFF 图片数据格式,与第 1 类数据样本的数据预设格式相同,则无须进行转化;如果第 1 类该测试数据的数据格式是 2576 像素×1932 像素 JPEG 图片数据格式,与第 1 类数据样本的数据预设格式不同,则须转化为 480 像素×640 像素的 TIFF 图片数据格式。

然后将该测试数据作为该类对应的深度学习模型的输入,通过该深度学习模型的计算得到该类对应的测试输出标签。例如,输入第 1 类任一测试数据“头像 1p”到第 1 类对应的深度学习模型,得到测试输出标签“张三身份证号”;输入第 2 类任一测试数据“语音 2q”到第 2 类对应的深度学习模型,得到测试输出标签“李四身份证号”。

步骤 S3-4,在步骤 S1 预处理的标签集中查找每一类的测试输出标签所在的标签集,之后判断该标签集是否只有一个标签元素;如果每一类的测试输出标签所在的标签集只有一个标签元素,则将每一类的测试输出标签作为该类的最优输出标签,即将步骤 S1-4 中“张三身份证号”“李四身份证号”分别作为第 1 类、第 2 类的最优输出标签;否则执行下一步。

步骤 S3-5,计算每一类的测试输出标签所对应的数据样本集与该类测试输出标签所在的标签集中每一标签元素对应的数据样本集的相似度,并根据该相似度计算并确定每一组可能输出标签;其中,每一组可能输出标签中包含了每一类的一个可能输出标签。例如,步

骤 S1-4 中“张三身份证号”所在第 1 类输出标签集为{朱一身份证号，郑二身份证号，张三身份证号，吴七身份证号}，对应的数据样本集为{朱一头像集，郑二头像集，张三头像集，吴七头像集}；“李四身份证号”所在第 2 类输出标签集为{田一身份证号，李四身份证号，吴七身份证号}，对应的数据样本集为{田一语音集，李四语音集，吴七语音集}。假设有 N 类数据样本集及对应的标签集，具体过程分为以下两种情况。

第一种情况：当 $N=1$ 时，只有一类数据样本集及对应的标签集。例如，仅存在上述第 1 类数据样本集及对应的标签集。

计算测试输出标签所对应的数据样本集{张三头像集}与测试输出标签所在的标签集中每一标签元素对应的数据样本集的相似度，则得到{张三头像集}与{朱一头像集}的相似度 a_1 为 80%，{张三头像集}与{郑二头像集}的相似度 a_2 为 90%，{张三头像集}与{张三头像集}的相似度 a_3 为 100%，{张三头像集}与{吴七头像集}的相似度 a_4 为 92%。其中 a_2、a_3、a_4 均大于第一预设值 80%，所以存在三组可能输出标签，分别为“郑二身份证号”“张三身份证号”“吴七身份证号”。

第二种情况：当 $N>1$ 时，有多类数据样本集及对应的标签集。例如 $N=2$，包括上述第 1 类数据样本集及对应的标签集、第 2 类数据样本集及对应的标签集。

首先计算第 1 类测试输出标签所对应的数据样本集{张三头像集}与测试输出标签所在的标签集中每一标签元素对应的数据样本集的相似度。相似度计算结果同上述 $N=1$ 的情况。

然后计算第 2 类测试输出标签所对应的数据样本集{李四语音集}与测试输出标签所在的标签集中每一标签元素对应的数据样本集的相似度。则得到{李四语音集}与{田一语音集}的相似度 b_1 为 95%，{李四语音集}与{李四语音集}的相似度 b_2 为 100%，{李四语音集}与{吴七语音集}的相似度 b_3 为 85%。

最后计算所有可能输出标签组各自的相似度第一综合值，输出标签组 c_1{朱一身份证号，田一身份证号}的相似度第一综合值 f_1 为 80%×95%=76%，输出标签组 c_2{朱一身份证号，李四身份证号}的相似度第一综合值 f_2 为 80%×100%=80%，输出标签组 c_3{朱一身份证号，吴七身份证号}的相似度第一综合值 f_3 为 80%×85%=68%，输出标签组 c_4{郑二身份证号，田一身份证号}的相似度第一综合值 f_4 为 90%×95%=85.5%，输出标签组 c_5{郑二身份证号，李四身份证号}的相似度第一综合值 f_5 为 90%×100%=90%，输出标签组 c_6{郑二身份证号，吴七身份证号}的相似度第一综合值 f_6 为 90%×85%=76.5%，输出标签组 c_7{张三身份证号，田一身份证号}的相似度第一综合值 f_7 为 100%×95%=95%，输出标签组 c_8{张三身份证号，李四身份证号}的相似度第一综合值 f_8 为 100%×100%=100%，输出标签组 c_9{张三身份证号，吴七身份证号}的相似度第一综合值 f_9 为 100%×85%=85%，输出标签组 c_{10}{吴七身份证号，田一身份证号}的相似度第一综合值 f_{10} 为 92%×95%=87.4%，输出标签组 c_{11}{吴七身份证号，李四身份证号}的相似度第一综合值 f_{11} 为 92%×100%=92%，输出标签组 c_{12}{吴七身份证号，吴七身份证号}的相似度第一综合值 p_{12} 为 92%×85%=78.2%。其中，f_4、f_5、f_7、f_8、f_{10}、f_{11} 均大于第二预设值 85%，所以有 6 组可能输出标签，分别为 c_4、c_5、c_7、c_8、c_{10}、c_{11} 对应的输出标签组。

（4）步骤 S4，计算由步骤 S3 得到的每一组可能输出标签中每一类的可能输出标签对应的数据样本集与该类测试数据集的相似度，并根据该相似度计算并确定一组可能输出标

签作为最优输出标签。对应步骤 S3 中的内容，具体过程分为以下两种情况。

第一种情况：当 $N=1$ 时，由步骤 S3 可知，有 3 组可能输出标签分别为“郑二身份证号”“张三身份证号”和“吴七身份证号”。其中可能输出标签“张三身份证号”的相似度值最大，因此将其作为最优输出标签组。

第二种情况：当 $N>1$ 时，由步骤 S3 可知，有 c_4、c_5、c_7、c_8、c_{10}、c_{11} 共 6 组可能输出标签，具体过程如下：

首先计算每一组可能输出标签中第 1 类的可能输出标签对应的数据样本集与该类测试数据集{张三头像集}的相似度，则得到{郑二头像集}与{张三头像集}的相似度为 90%，{张三头像集}与{张三头像集}的相似度为 100%，{吴七头像集}与{张三头像集}的相似度为 92%。

然后计算每一组可能输出标签中第 2 类的可能输出标签对应的数据样本集与该类测试数据集{李四语音集}的相似度，则得到{李四语音集}与{李四语音集}的相似度为 100%，{田一语音集}与{李四语音集}的相似度为 95%。

则 c_4 输出标签组{郑二身份证号，田一身份证号}相似度第二综合值 g_4 为 90%×95%=85.5%，c_5 输出标签组{郑二身份证号，李四身份证号}相似度第二综合值 g_5 为 90%×100%=90%，c_7 输出标签组{张三身份证号，田一身份证号}相似度第二综合值 g_7 为 100%×95%=95%，c_8 输出标签组{张三身份证号，李四身份证号}相似度第二综合值 g_8 为 100%×100%=100%，c_{10} 输出标签组{吴七身份证号，田一身份证号}相似度第二综合值 g_{10} 为 92%×95%=87.4%，c_{11} 输出标签组{吴七身份证号，李四身份证号}相似度第二综合值 g_{11} 为 92%×100%=92%。其中相似度第二综合值最大的是 g_8，因此将 c_8 输出标签组{张三身份证号，李四身份证号}作为最优输出标签组。

(5) 步骤 S5，计算可能输出标签中各类输出标签都一致及不一致的概率，作为各类输出标签一致及不一致的概率。对应步骤 S3 中的内容，具体过程分为以下两种情况。

第一种情况：当 $N=1$ 时，由步骤 S3 可知，只有一类输出标签，因此各类输出标签一致的概率为 100%，不一致的概率为 0。

第二种情况：当 $N>1$ 时，由步骤 S3 可知，有 c_4、c_5、c_7、c_8、c_{10}、c_{11} 共 6 组可能输出标签。其中 c_4 输出标签组{郑二身份证号，田一身份证号}中各类可能输出标签不一致，c_5 输出标签组{郑二身份证号，李四身份证号}中各类可能输出标签不一致，c_7 输出标签组{张三身份证号，田一身份证号}中各类可能输出标签不一致，c_8 输出标签组{张三身份证号，李四身份证号}中各类可能输出标签不一致，c_{10} 输出标签组{吴七身份证号，田一身份证号}中各类可能输出标签不一致，c_{11} 输出标签组{吴七身份证号，李四身份证号}中各类可能输出标签不一致。由上述可知，所有可能输出标签组中各类可能输出标签均不一致，所以各类输出标签一致的概率为 0，不一致的概率为 100%，说明各类测试样本对应同一个人的概率为 0。

为了进一步阐明过程，假设有 4 组可能输出标签 d_4、d_6、d_{10}、d_{11}，d_4 输出标签组{郑二身份证号，郑二身份证号}中各类可能输出标签一致，对应的相似度第二综合值为 89%；d_6 输出标签组{张三身份证号，张三身份证号}中各类可能输出标签一致，对应的相似度第二综合值为 53%；d_{10} 输出标签组{郑二身份证号，李四身份证号}中各类可能输出标签不一致，对应的相似度第二综合值为 67%；d_{11} 输出标签组{张三身份证号，郑二身份证号}中各类可能输出标签不一致，对应的相似度第二综合值为 75%。

判断结果为一致的各组可能输出标签对应的相似度第二综合值之和(89%+53%)除以所有可能输出标签对应的相似度第二综合值之和(89%+53%+67%+75%),得到各类输出标签一致的概率为50%。以100%减去各类输出标签一致的概率50%,则得到各类输出标签不一致的概率为50%。

(6)步骤S6,将可能输出标签、最优输出标签、各类输出标签一致及不一致的概率输出。

1.2.2　联合聚类深度学习模型的数据识别系统

联合聚类深度学习模型的数据识别系统包括:

(1)采集模块1,用于 N 类数据样本集及每一类数据样本集对应的标签集,再获取 N 类数据样本集中每一类数据样本的数据预设格式,并获取标签预设格式,然后对 N 类数据样本集和标签集进行预处理;$N \geq 1$。

(2)初始化模块2,用于初始化 N 类数据样本集对应的 N 个深度学习模型。

(3)训练模块3,用于将获取的每一类数据样本集作为输入,将其对应的标签集作为输出,对该类对应的深度学习模型进行训练,得到 N 个训练后的深度学习模型。

(4)测试模块4,用于为每一类对应的深度学习模型获取一个测试数据,将每一类测试数据的数据格式转化为该类数据样本的数据预设格式,然后将该测试数据作为该类对应的深度学习模型的输入,通过该深度学习模型的计算得到该类对应的测试输出标签。

(5)判别模块5,用于在预处理的标签集中查找每一类的测试输出标签所在的标签集,之后判断该标签集是否只有一个标签元素。如果每一类的测试输出标签所在的标签集只有一个标签元素,则将每一类的测试输出标签作为该类的最优输出标签;否则执行下一步。

(6)第一相似度计算模块6,用于计算每一类的测试输出标签所对应的数据样本集与该类测试输出标签所在的标签集中每一标签元素对应的数据样本集的相似度,根据该相似度计算并确定每一组可能输出标签;其中,每一组可能输出标签中包含了每一类的一个可能输出标签。

(7)第二相似度计算模块7,用于计算每一组可能输出标签中每一类的可能输出标签对应的数据样本集与该类测试数据集的相似度,并根据该相似度计算并确定一组可能输出标签作为最优输出标签。

(8)概率统计模块8,用于计算可能输出标签中各类输出标签都一致及不一致的概率,作为各类输出标签一致及不一致的概率。

(9)输出模块9,用于将可能输出标签、最优输出标签、各类输出标签一致及不一致的概率输出。

1.3　基于深度学习模型的身份识别

现有技术中深度学习技术能通过输入数据得到输出标签(例如,通过头像得到该人身份证号,或者通过语音得到该人身份证号),在自顶向下的监督训练阶段需要通过带有标签的

数据(例如,带有身份证号的头像,或者带有身份证号的语音)进行监督训练(即将数据样本作为深度学习模型的输入,标签作为深度学习模型的输出)。

但是,由于深度学习模型的准确率无法达到100%,所以一个头像输入深度学习模型后输出的身份证号可能是另一个头像相似的人的身份证号,一个语音输入深度学习模型后输出的身份证号可能是另一个语音相似的人的身份证号。进行用户身份识别时,为了提高识别的准确性,一般需要采集用户的多类数据,进行综合判断,但这样会增加采集的成本,而且采集越多类数据就会给用户造成越多的不便,并且在很大程度上耽误了用户的时间,所以需要通过尽量少的测试类别来得到准确的判断。但现有深度学习模型技术无法实现通过多类深度学习模型进行综合辨别[5,6],也无法进行类别的选择。

本技术研究并给出了一种新的基于深度学习模型的身份识别方法和系统,具体为:①获取 N 类输入数据;②初始化对应的 N 类深度学习模型;③训练 N 类深度学习模型;④根据采集成本从低到高对类别进行排序;⑤初始化 i、最优输出标签 L、最大相似度相对比值 U;⑥计算 T_i 类测试输出标签 L_i,并判断 L 是否与 L_i 相同;⑦获取 T_i 类相似度相对比值 U_i,并根据上一步的结果判别是身份识别失败还是执行下一步;⑧更新 L 和 U;⑨根据 U 与预设最大相似度相对比值 c 的关系、i 与 N 的关系,判别身份是否识别成功、是否需要将 i 加1继续循环执行。本技术方案优先选择低成本测试数据,并逐渐增加测试输入类型,在最低成本的情况下取得最高的判别准确率。

1.3.1 基于深度学习模型的身份识别方法

1. 基于深度学习模型的 *N* 类身份识别

一种基于深度学习模型的身份识别方法,包括以下步骤。

(1) 步骤S1,获取用于身份识别的 N 类输入数据。

(2) 步骤S2,初始化 N 类输入数据对应的 N 类深度学习模型。具体为:

步骤S2-1,初始化每一类深度学习模型的输入格式为对应类的输入数据的格式。

步骤S2-2,初始化每一类深度学习模型的输出格式为用户身份信息的格式。

(3) 步骤S3,对 N 类深度学习模型进行训练。具体为:

步骤S3-1,获取 N 类中每一类训练输入数据,并将其作为该类对应的深度学习模型的输入,进行自下而上的无监督训练。

步骤S3-2,获取 N 类中每一类带有标签的训练输入数据及其对应的预期输出标签,并将其分别作为该类对应的深度学习模型的输入数据与预期输出标签,进行自顶向下的有监督训练。

(4) 步骤S4,获取 N 类输入数据中每一类输入数据的采集成本,根据采集成本从低到高对类别进行排序,得到 N 类已排序的类别为 $T_1,T_2,\cdots,T_N$;其中采集成本是经济成本与时间成本的加权平均。

(5) 步骤S5,初始化 $i=1$,待识别身份用户的最优输出标签 $L=\varnothing$,最大相似度相对比值 $U=0$;其中 i 为自然数,取值范围为 $1\leqslant i\leqslant N$。

(6) 步骤S6,获取一个待识别身份用户的 T_i 类测试输入数据,将其作为 T_i 类深度学习

模型的输入，通过 T_i 类深度学习模型计算得到 T_i 类测试输出标签 L_i；判断最优输出标签 L 是否为空或与 T_i 类测试输出标签 L_i 相同，如果为空或相同，则执行步骤 S7，否则执行步骤 S8。

(7) 步骤 S7，获取 T_i 类相似度相对比值 U_i，然后执行步骤 S9。

其中，获取 T_i 类相似度相对比值 U_i 具体步骤为：

首先从 T_i 类训练输入数据集中搜索预期标签与 T_i 类测试输出标签 L_i 匹配的 T_i 类训练输入数据集，将其作为 T_i 类第一训练输入数据集；然后计算 T_i 类第一训练输入数据集中每一训练输入数据与 T_i 类测试输入数据的相似度，将其中最大的相似度 M_i 与 T_i 类测试相似度预设阈值 a_i 的比值作为 T_i 类相似度相对比值 U_i。

(8) 步骤 S8，获取 T_i 类相似度相对比值 U_i，并判别是身份识别失败还是执行步骤 S9。

其中，获取 T_i 类相似度相对比值 U_i 具体步骤为：

首先从 T_i 类训练输入数据集中搜索预期标签与 T_i 类测试输出标签 L_i 匹配的 T_i 类训练输入数据集，将其作为 T_i 类第一训练输入数据集；然后计算 T_i 类第一训练输入数据集中每一训练输入数据与 T_i 类测试输入数据的相似度，将其中最大的相似度 M_i 与 T_i 类测试相似度预设阈值 a_i 的比值作为 T_i 类相似度相对比值 U_i。

其中，判别是身份识别失败还是执行步骤 S9，具体为步骤：

步骤 S8-1，从 T_i 类训练输入数据集中搜索预期标签与 T_i 类测试输出标签 L_i 匹配的 T_i 类训练输入数据集，将其作为 T_i 类第一训练输入数据集。

步骤 S8-2，从 T_i 类训练输入数据集中搜索预期标签与最优输出标签 L 匹配的 T_i 类训练输入数据集，将其作为 T_i 类第二训练输入数据集。

步骤 S8-3，计算 T_i 类第一训练输入数据集中每一训练输入数据与第二训练输入数据集中每一训练输入数据的相似度，如果其中最大的相似度 N_i 大于等于 T_i 类混淆相似度预设阈值 b_i，则执行步骤 S9，否则为身份识别失败。

(9) 步骤 S9，判别 T_i 类相似度相对比值 U_i 与最大相似度相对比值 U 之间的关系，更新最优输出标签 L 和最大相似度相对比值 U。具体步骤为：

如果 $U_i > U$，则将最优输出标签 L 设置为 T_i 类输出标签，将最大相似度相对比值 U 设置为 T_i 类相似度相对比值 U_i；否则保持最优输出标签 L 和最大相似度相对比值 U 不变。

(10) 步骤 S10，判别最大相似度相对比值 U 与预设最大相似度相对比值 c 之间的关系；如果 $U > c$，则身份识别成功，将最优输出标签作为识别出的身份标签；否则判别 i 与 N 之间的关系，如果 $i \geqslant N$，则身份识别失败，如果 $i < N$，则将 i 加 1 并返回步骤 S6 继续执行。

2. 基于深度学习模型的 3 类身份识别

结合图 1-7，以 $N=3$ 为例，本技术方案基于深度学习模型的身份识别方法，包括以下步骤：

(1) 步骤 S1，获取用于身份识别的 3 类输入数据，分别为图像类、指纹类和语音类输入数据。

(2) 步骤 S2，初始化 3 类输入数据对应的 3 类深度学习模型。具体步骤为：

步骤 S2-1，初始化每一类深度学习模型的输入格式为对应类输入数据的格式。

例如，初始化图像类深度学习模型的输入格式为图像类输入数据的格式。

步骤 S2-2，初始化每一类深度学习模型的输出格式为用户身份信息的格式。例如，初始化图像类深度学习模型的输出格式为用户身份信息的格式，如身份证号。

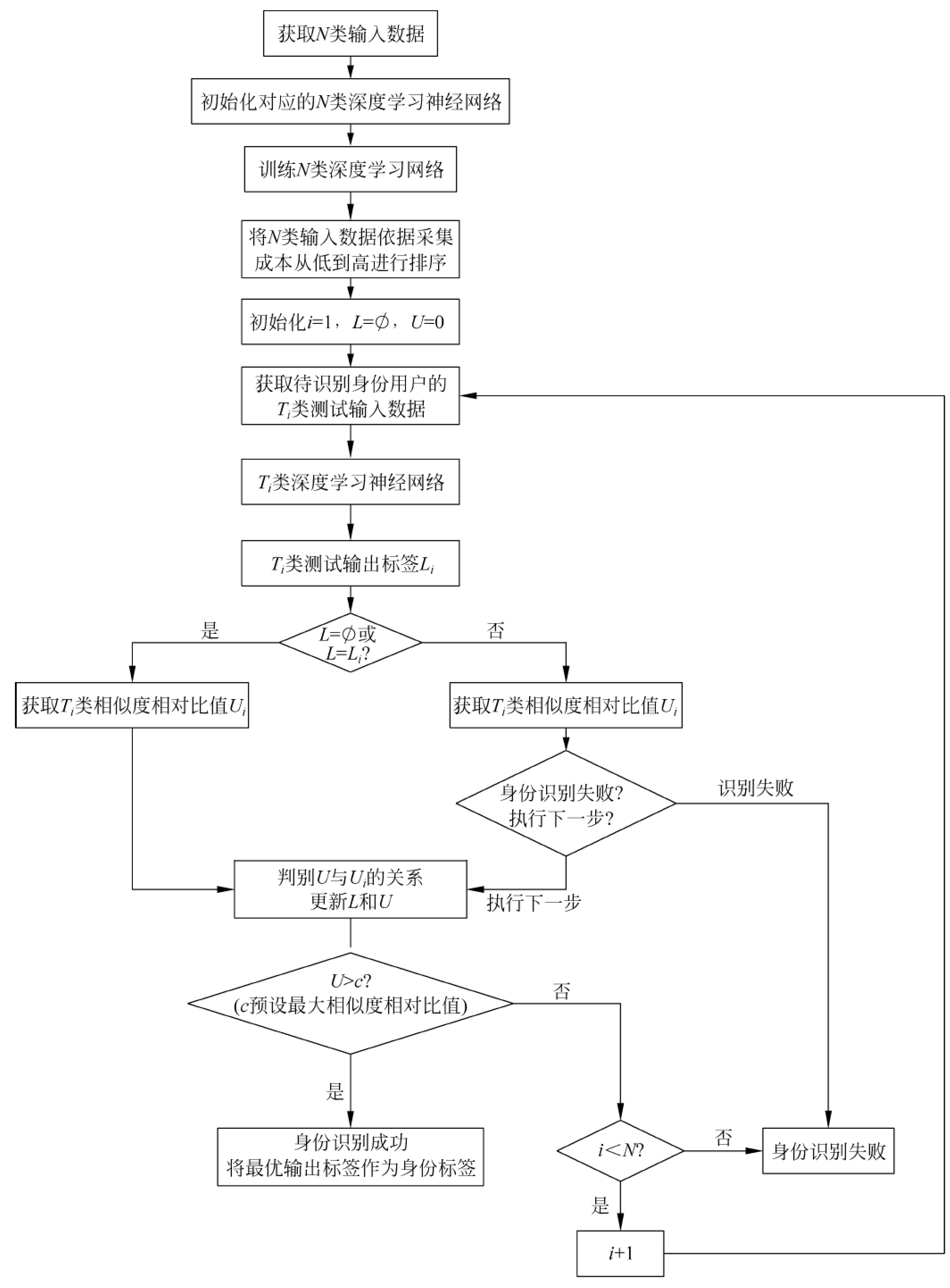


图 1-7　基于深度学习模型的身份识别方法流程

（3）步骤 S3，对 3 类深度学习模型进行训练。结合图 1-8，具体步骤为：

步骤 S3-1，获取 3 类中每一类训练输入数据，并将每一类训练输入数据作为该类对应的深度学习模型的输入，进行自下而上的无监督训练。例如，将每一个头像作为头像类深度学习模型的输入，进行自下而上的无监督训练。

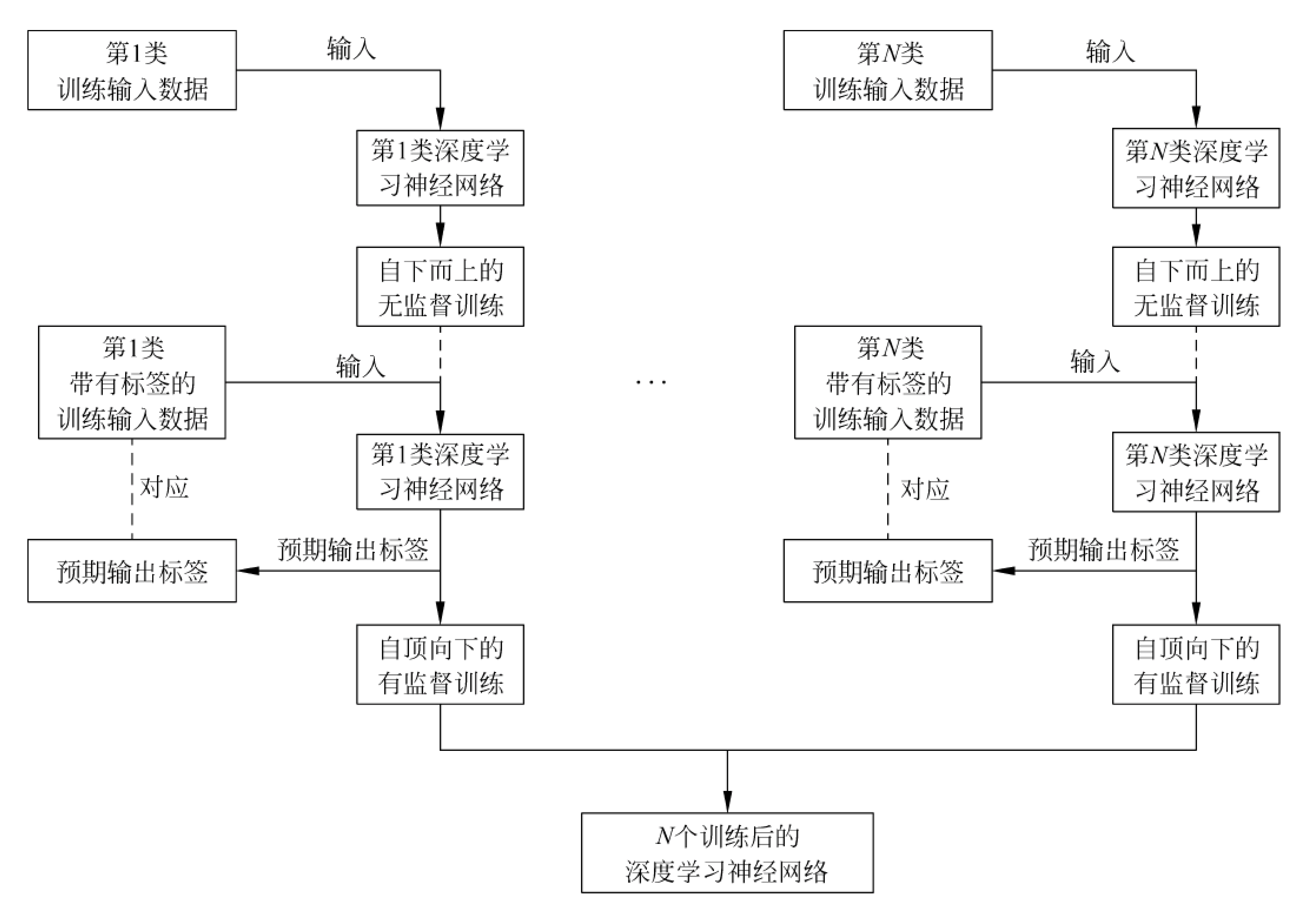


图 1-8 基于深度学习模型的身份识别方法中对深度学习模型进行训练流程

步骤 S3-2，获取 3 类中每一类带有标签的训练输入数据及其对应的预期输出标签，并将其分别作为该类对应的深度学习模型的输入数据与预期输出标签，进行自顶向下的有监督训练。例如，将带有姓名标签的头像及其对应的姓名标签分别作为头像类深度学习模型的输入数据与预期输出标签，进行自顶向下的监督训练。

（4）步骤 S4，获取 3 类输入数据中每一类输入数据的采集成本，根据采集成本从低到高对类别进行排序，得到 3 类已排序的类别为 T_1，T_2，T_3；其中采集成本是经济成本与时间成本的加权平均。例如，图像类采集的经济成本是 1 元，时间是 0.3min，则采集成本为加权平均(1+0.3)/2=0.65；指纹类采集的经济成本是 4 元，时间是 1min，则采集成本为加权平均(4+1)/2=2.5；语音类采集的经济成本是 2 元，时间是 2min，则采集成本为加权平均(2+2)/2=2。根据采集成本从低到高对类别进行排序，得到 3 类已排序的类别 T_1 图像类、T_2 语音类、T_3 指纹类。

（5）步骤 S5，初始化 $i=1$，待识别身份用户的最优输出标签 $L=\varnothing$，最大相似度相对比值 $U=0$；其中 i 为自然数，取值范围为 $1\leqslant i\leqslant N$。

（6）步骤 S6，获取一个待识别身份用户的 T_1 图像类测试输入数据，将其作为 T_1 图像类深度学习模型的输入，通过 T_1 图像类深度学习模型计算得到 T_1 类测试输出标签 L_1，例如“张三身份证号”，而当前最优输出标签 $L=\varnothing$，则执行步骤 S7。

（7）步骤 S7，获取 T_1 头像类相似度相对比值 U_1。从头像类训练输入数据集中搜索出预期输出标签与头像类测试输出标签 L_1“张三身份证号”匹配的头像类训练输入数据集，作为头像类第一训练输入数据集；计算头像类第一训练输入数据集中每一训练输入数据与头

像类测试输入数据之间的相似度(例如,分别为 70%、86%、44%、24%、78%、98%),将其中最大相似度 98%与头像类相似度预设阈值(例如,为 82%)的比值作为 T_1 头像类相似度相对比值,则 $U_1=1.19$。

(8) 步骤 S8,因 $U_1>U$,则将最优输出标签 L 设置为 T_1 头像类输出标签“张三身份证号”,将最大相似度相对比值 U 设置为 T_1 图像类相似度相对比值 U_1,即 $U=1.19$。

(9) 步骤 S9,假设预设最大相似度相对比值 $c=1.2$,因 $U<c$,则判别 i 与 N 之间的关系,由 $i=1$、$N=3$ 得 $i<N$,则将 i 加 1 并返回步骤 S6 继续执行。具体步骤为:

由上可知 $i=2$,获取一个待识别身份用户的 T_2 语音类测试输入数据,将其作为 T_2 语音类深度学习模型的输入,通过 T_2 语音类深度学习模型计算得到 T_2 类测试输出标签 L_2,例如“张三身份证号”,而当前最优输出标签 L 为“张三身份证号”,L 与 L_2 相同,则执行如下步骤:

获取 T_2 语音类相似度相对比值 U_2。从语音类训练输入数据集中搜索出预期输出标签与语音类测试输出标签 L_2“张三身份证号”匹配的语音类训练输入数据集,作为语音类第一训练输入数据集;计算语音类第一训练输入数据集中每一训练输入数据与语音类测试输入数据之间的相似度(例如,分别为 52%、64%、78%、74%、68%),将其中最大相似度 78%与语音类相似度预设阈值(例如,为 78%)的比值作为 T_2 语音类相似度相对比值,则 $U_2=1$。

因 $U_2<U$,则保持最优输出标签 L“张三身份证号”、最大相似度相对比值 $U=1.19$ 不变。

因 $U<c$,则判别 i 与 N 之间的关系,由 $i=2$、$N=3$ 得 $i<N$,则将 i 加 1 并返回步骤 S6 继续执行。具体步骤为:

由上可知 $i=3$,获取一个待识别身份用户的 T_3 指纹类测试输入数据,将其作为 T_3 指纹类深度学习模型的输入,通过 T_3 指纹类深度学习模型计算得到 T_3 类测试输出标签 L_3,例如“李四身份证号”,而当前最优输出标签 L 为“张三身份证号”,L 与 L_3 不相同,则执行如下步骤:

首先获取 T_3 指纹类相似度相对比值 U_3。从指纹类训练输入数据集中搜索出预期输出标签与指纹类测试输出标签 L_3“李四身份证号”匹配的指纹类训练输入数据集,作为指纹类第一训练输入数据集,计算其中每一训练输入数据与指纹类测试输入数据之间的相似度(例如为 87%),将其中最大相似度 87%与指纹类相似度预设阈值(例如为 70%)的比值作为 T_3 指纹类相似度相对比值,则 $U_3=1.24$。

然后判断是身份识别失败还是执行步骤 S8,见图 1-9。从指纹类训练输入数据集中搜索出预期输出标签与当前最优输出标签“张三身份证号”匹配的指纹类训练输入数据集,作为指纹类第二训练输入数据集,计算其中每一训练输入数据与第一训练输入数据集中每一训练输入数据之间的相似度(例如,分别为 46%、73%、58%、……、89%、62%),如果其中最大相似度 89%小于指纹类混淆相似度预设阈值,则识别验证失败(例如,如果指纹类混淆相似度预设阈值是 90%,那么张三与李四对应指纹类训练输入数据集的最大相似度 89%小于指纹类混淆相似度预设阈值 90%,说明张三的指纹与李四的不够相似,而测试输出的结果是张三、李四,这是矛盾的,那么就可能是用户在伪造输入数据,所以识别验证失败);如果其中最大相似度大于等于指纹类混淆相似度预设阈值,则执行步骤 S8(例如,如果指纹类混淆相似度预设阈值是 88%,那么张三与李四对应指纹类训练输入数据集的最大相似度 89%小于指纹类混淆相似度预设阈值 88%,说明张三的指纹与李四的足够相似,而测试输出的

结果是张三、李四，有可能是因为两者相似导致的误判，所以需要继续增加其他类的输入数据进一步识别)。假设上述判断结果为执行步骤 S8，则执行如下：因 $U_3>U$，则将最优输出标签 L 设置为 T_3 指纹类输出标签“李四身份证号”，将最大相似度相对比值 U 设置为 T_3 指纹类相似度相对比值 U_3，即 $U=1.24$。

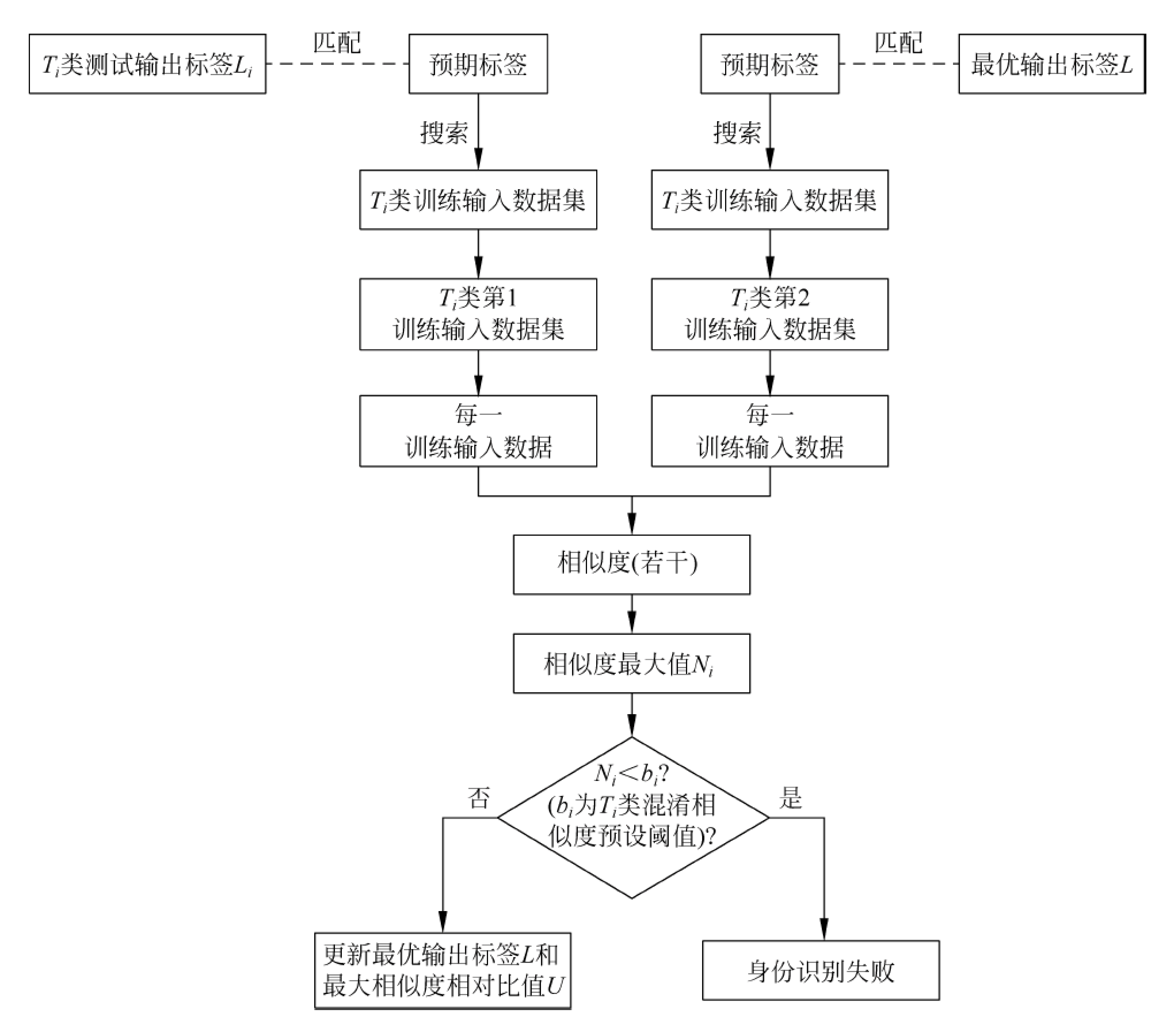


图 1-9　基于深度学习模型的身份识别方法中判别身份识别失败还是更新最优输出标签和最大相似度相对比值的流程

因为 $U>c$，则身份识别成功，将最优输出标签“李四身份证号”作为识别出的身份标签。

综上技术方案可知，深度学习模型已经事先通过训练输入数据训练好了，所以在身份识别时只需要输入测试输入数据，就能通过深度学习模型实时计算得到输出标签，从而运算速度要比传统身份识别方式“将测试输入数据与所有海量(例如，1000 个用户身份可能对应 10 万个样本)训练数据一一匹配(需要匹配 10 万次)，将最匹配的训练数据对应的标签作为输出标签”快得多。但深度学习模型的输出标签也不一定 100%准确，此时在训练数据中搜索出该深度学习模型的输出标签对应的训练数据(根据标签检索速度非常快，而一个输出标签对应的一个用户的训练数据比较少，例如 100 个左右)与测试数据进行比较，来判断深度学习模型的输出标签是否准确。这种方式计算速度快，又结合了深度学习模型和传统模式匹配的各自优势，使得深度学习模型输出标签的准确性有了评价的标准。

不同类别输入数据相似度的阈值标准是不同的，例如图像之间的相似度预设阈值是 90%，达到这个相似度阈值就表明这两个图像非常像，可以认为是可能对应同一个身份标签

了；而语音之间的相似度预设阈值是80%即可，因为语音的噪声相对大些，达到80%的相似度预设阈值就表明这两个语音非常像，可以认为是可能对应同一个身份标签了。

而将每一类最大相似度与该类相似度预设阈值的比值作为该类最大相似度相对比值之后，不同类的最大相似度相对比值就可以相互比较了，因为已经通过除以该类相似度预设阈值进行归一化了。例如，图像最大相似度相对比值与语音最大相似度相对比值就可以比较，越大表明其对应的输出身份标签越准确。所以如果当前最大相似度相对比值大于预设最大相似度相对比值（例如1.2），则身份识别就成功，那么就无须继续输入其他类别的测试输入数据了，节省了测试的成本和用户的时间，使得身份识别更为低成本且高效。

1.3.2 基于深度学习模型的身份识别系统

一种基于深度学习模型的身份识别系统如图1-10所示，包括输入数据获取模块、初始化神经网络模块、训练模块、排序模块、初始化起始识别条件模块、测试输出标签获取模块、最优输出标签判别模块、相似度相对比值获取及判别模块、数据更新模块、身份识别判定模块。

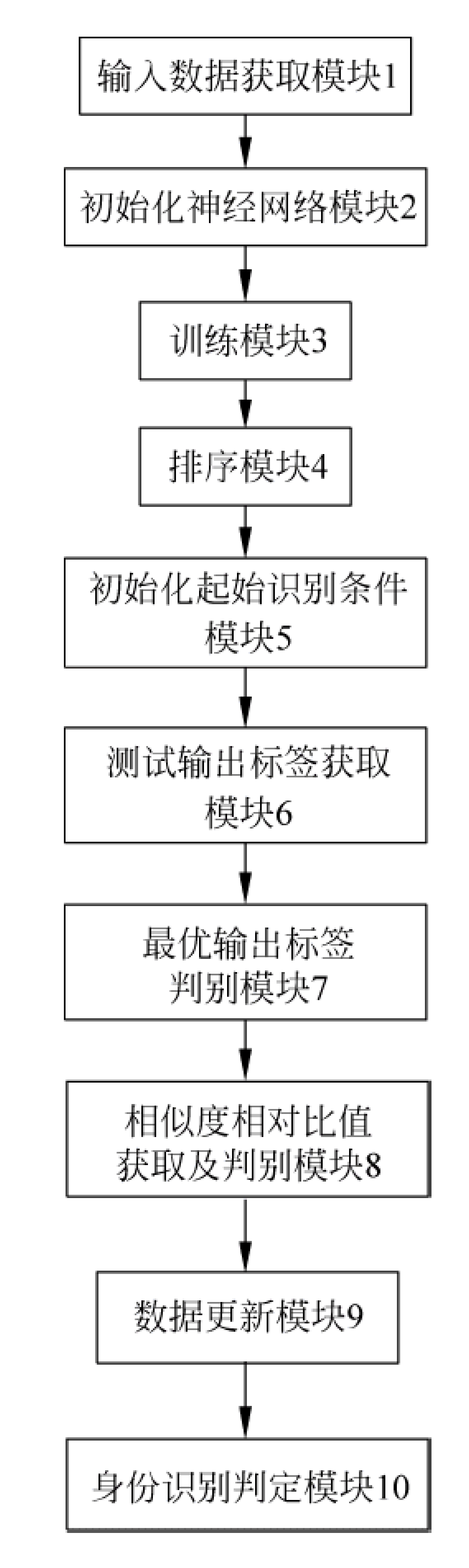


图1-10 基于深度学习模型的身份识别系统结构框图

(1) 输入数据获取模块1，获取用于身份识别的 N 类输入数据。

(2) 初始化神经网络模块2，如图1-11所示，用于初始化 N 类输入数据对应的 N 类深度学习模型。具体包括：

① 初始化输入格式单元2-1，用于初始化每一类深度学习模型的输入格式为对应类的输入数据的格式；

② 初始化输出格式单元2-2，用于初始化每一类深度学习模型的输出格式为用户身份信息的格式。

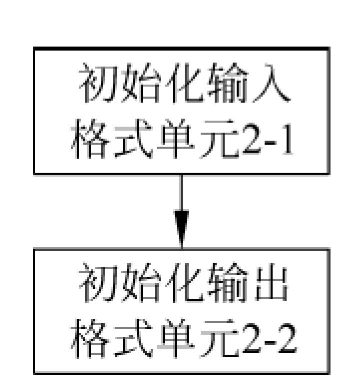
(3) 训练模块3，如图1-12所示，用于对 N 类深度学习模型进行训练。具体包括：

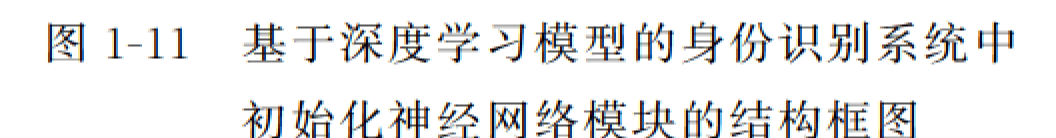
① 无监督训练单元3-1，用于获取 N 类中每一类训练输入数据，并将每一类训练输入数据作为该类对应的深度学习模型的输入，进行自下而上的无监督训练；

② 有监督训练单元3-2，用于获取 N 类中每一类带有标签的训练输入数据及其对应的预期输出标签，并将其分别作为该类对应的深度学习模型的输入数据与预期输出标签，进行自顶向下的有监督训练。

(4) 排序模块4，用于将 N 类输入数据依据采集成本从低到高进行排序。

(5) 初始化起始识别条件模块5，用于初始化 $i=1$，待识别身份用户的最优输出标签 $L=\varnothing$，最大相似度相对比值 $U=0$。




图 1-11 基于深度学习模型的身份识别系统中初始化神经网络模块的结构框图

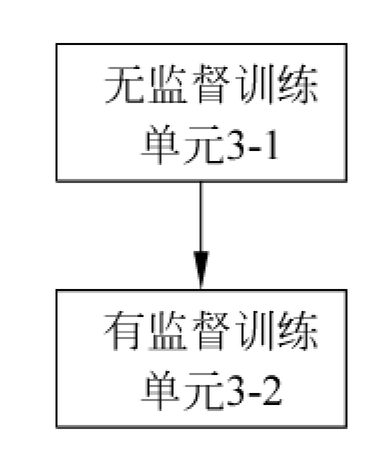


图 1-12 基于深度学习模型的身份识别系统中训练模块的结构框图

（6）测试输出标签获取模块 6，用于将一个待识别身份用户的 T_i 类测试输入数据作为 T_i 类深度学习模型的输入，通过 T_i 类深度学习模型计算得到 T_i 类测试输出标签 L_i。

（7）最优输出标签判别模块 7，用于判断最优输出标签 L 是否为空或与 T_i 类测试输出标签 L_i 相同。

（8）相似度相对比值获取及判别模块 8 如图 1-13 所示，用于获取 T_i 类相似度相对比值 U_i，并判别直接执行下一模块还是需要进一步判别是否身份识别失败后再执行下一模块。具体包括：

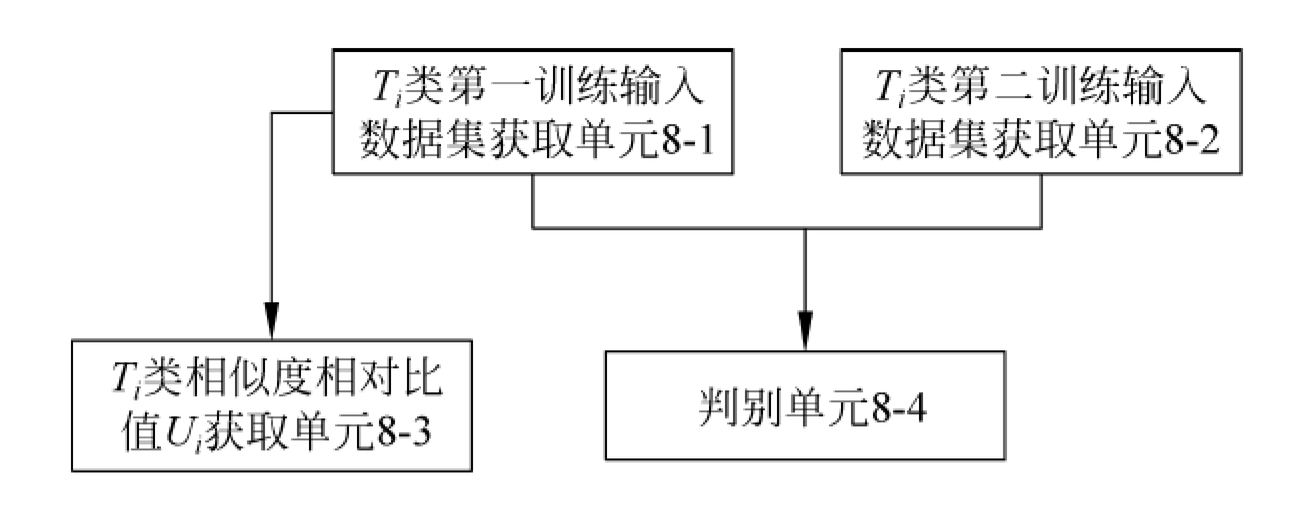


图 1-13 基于深度学习模型的身份识别系统中相似度相对比值获取及判别模块的结构框图

① T_i 类第一训练输入数据集获取单元 8-1，用于从 T_i 类训练输入数据集中搜索预期标签与 T_i 类测试输出标签 L_i 匹配的 T_i 类训练输入数据集，将其作为 T_i 类第一训练输入数据集；

② T_i 类第二训练输入数据集获取单元 8-2，用于从 T_i 类训练输入数据集中搜索预期标签与最优输出标签 L 匹配的 T_i 类训练输入数据集，将其作为 T_i 类第二训练输入数据集；

③ T_i 类相似度相对比值 U_i 获取单元 8-3，用于计算 T_i 类第一训练输入数据集中每一训练输入数据与 T_i 类测试输入数据的相似度，将其中最大的相似度 M_i 与 T_i 类测试相似度预设阈值 a_i 的比值作为 T_i 类相似度相对比值 U_i；

④ 判别单元 8-4，用于依据最优输出标签判别模块的结果，判别是直接执行下一模块还是需要进一步判别是否身份识别失败后再执行下一模块。

（9）数据更新模块 9，用于依据 T_i 类相似度相对比值 U_i 与最大相似度相对比值 U 之间的关系，更新最优输出标签 L 和最大相似度相对比值 U。

（10）身份识别判定模块 10，用于依据最大相似度相对比值 U 与预设最大相似度相对比值 c 之间的关系、i 与 N 之间的关系，判定身份识别是成功、失败还是 i 加 1 返回测试输出

标签获取模块继续执行。

本技术方案与现有技术相比，其显著优点为：本技术方案通过训练多类深度学习模型，在对用户进行身份识别时，优先选择低成本的测试数据。本技术方案先以一种类型的测试输入数据进行判别，如果判别不出用户的身份，则逐渐增加其他类型的测试输入数据进行判别，从而在最低成本的情况下取得最高的判别准确率。



第2章

智能型深度学习模型

将现有技术中深度学习模型的初始化环节或训练环节或整个环节进行优化，能够使得深度学习模型更为智能高效。本章给出了两种智能型深度学习模型：基于大数据的深度学习模型初始化、深度学习模型进化。基于大数据的深度学习模型初始化方法包括获取目标初始化深度学习模型的特征信息，将特征信息在预设深度学习模型数据库中进行匹配，得到匹配结果，根据匹配结果确定目标初始化深度学习模型，通过训练数据训练目标初始化深度学习模型，得到深度学习的目标深度学习模型。本技术方案通过特征信息匹配，将预先存储的以往训练过的深度学习模型作为将来深度学习模型初始化的基础，从而优化深度学习模型初始化的结果，使得深度学习模型在更优初始化的情况下进行训练，从而能够缩短训练时间，提高深度学习模型的训练速度，同时也能保证深度学习模型深度学习的效果。深度学习模型进化方法包括：将输入数据输入调取的已训练的深度学习模型得到输出数据，记录使用时间及获取对应的真实结果数据；按照预设的抽查规则选取使用过程中的输入数据、输出数据、使用时间和真实结果数据；将选取的输入数据及对应的输出数据、真实结果数据和使用时间组成数据组；计算数据组中输出数据与真实结果数据的差异值，统计预设时间段内所有数据组中差异值大于预设阈值的组数；若组数大于或等于存储的组数阈值，根据差异组的输入数据和真实结果数据对已训练的深度学习模型进行训练得到进化后的深度学习模型。如此可纠正深度学习模型的偏差，提高准确率。

2.1　基于大数据的深度学习模型初始化

深度学习模型作为对人脑最简单的一种抽象和模拟，是模仿人的大脑神经系统信息处理功能的一个人工智能化系统。深度学习模型是以数学和物理方法以及信息处理的角度对

人脑深度学习模型进行抽象，并建立某种简化模型，旨在模仿人脑结构及其功能的信息处理系统。深度学习模型最有吸引力的特点就是它的学习能力，其中，深度学习本质上是一种采用多层非线性变换的信息提取方法，通过无监督或有监督训练得出一组能够提取数据丰富内涵的网络参数，最终用于特征提取或转换以及模式分类等任务。

传统深度学习模型的初始化都是对深度学习模型随机地赋予初值或按照与待训练数据无关的规则赋予初值，然后进行学习训练，最终获得网络模型[7,8]。然而，通过随机赋予初值训练成符合训练样本数据的权值需要很长时间，效率较低，对于结构复杂、层数多的深度学习模型来说尤其如此。

本技术研究并给出了一种新的深度学习模型深度学习方法、系统、计算机可读存储介质和计算机设备。

2.1.1　基于大数据的深度学习模型初始化方法

1. 基于大数据的深度学习模型初始化的基本方法

深度学习模型深度学习方法的流程示意图如图 2-1 所示，本技术方案中的深度学习模型深度学习方法包括：

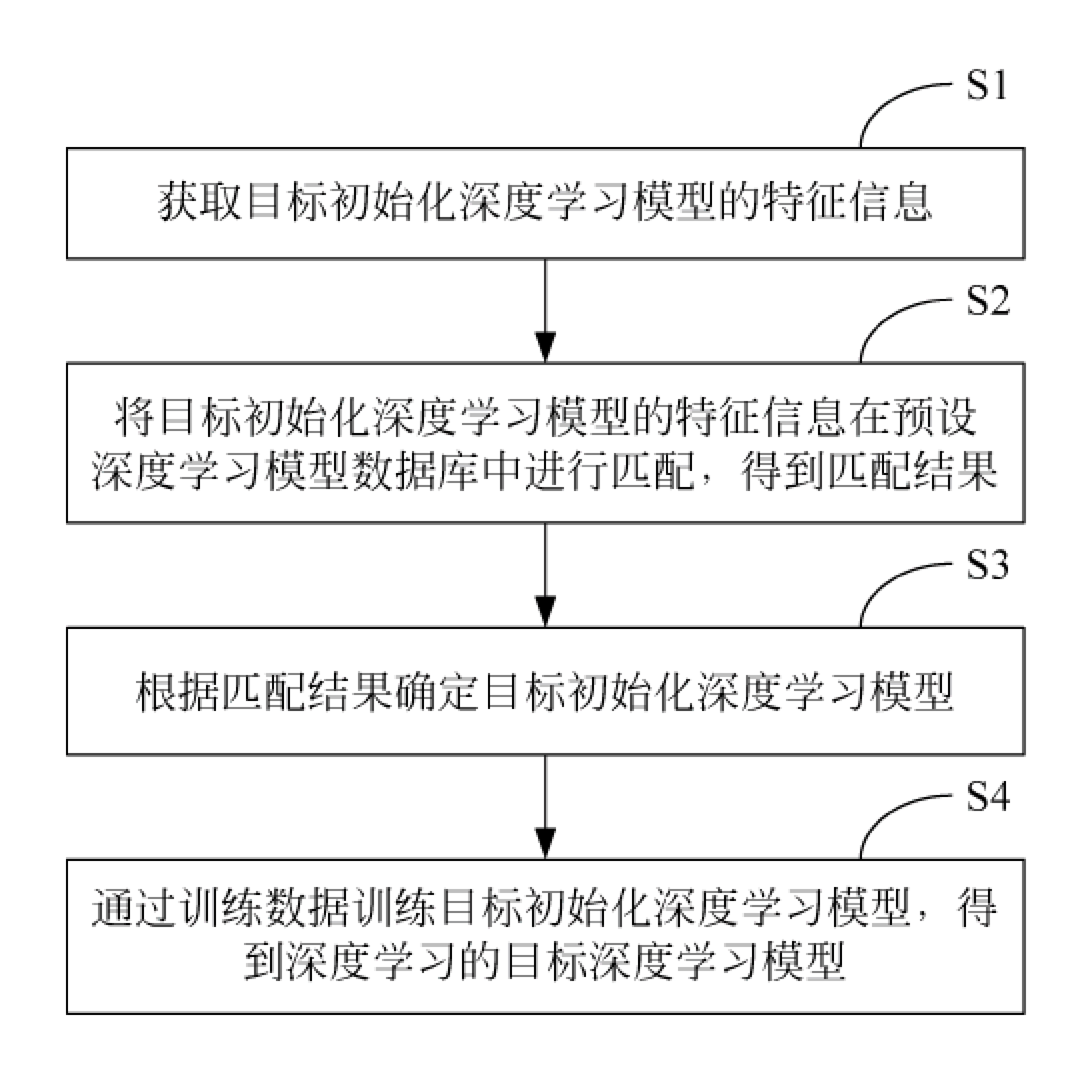


图 2-1　深度学习模型深度学习方法的流程示意图

（1）步骤 S1，获取目标初始化深度学习模型的特征信息。

具体步骤为：根据需要构建的目标初始化深度学习模型确定其特征信息，特征信息可以为描述该深度学习模型的参数，如目标初始化深度学习模型的应用信息，应用信息可以包括训练数据的格式，例如图片格式、是否类标签等；应用的功能，例如人脸识别、语音识别等。具体而言，例如，需要初始化一个用于人脸识别的深度学习模型，该深度学习模型训练数据的输入是图像数据，输出是标签“是人脸”“不是人脸”，那么此时该目标初始化深度学习模型的特征信息就是：训练数据的输入格式为“图片格式”，输出格式“是否类标签”，功能为“人脸识别”。

（2）步骤 S2，将目标初始化深度学习模型的特征信息在预设深度学习模型数据库中进行匹配，得到匹配结果。

在获得目标初始化深度学习模型的特征信息后，连接预设的深度学习模型数据库，该深度学习模型数据库包括预先采集的各种深度学习模型的大数据，可以包括已训练并实用的深度学习模型，还可以包括部分训练的深度学习模型。例如，可以包括用于人脸识别、指纹识别、汉语识别、英语识别等各种已训练的深度学习模型。相应的，深度学习模型数据库中的各种深度学习模型也设有其对应的特征信息。根据目标初始化深度学习模型的特征信息在预设深度学习模型数据库中与各深度学习模型对应的特征信息进行搜索匹配，得到匹配结果。在具体应用时，可以在匹配结果中选择匹配程度最高的特征信息对应的已训练深度学习模型，从而获得与目标初始化深度学习模型的特征信息相似度最高的最优已训练深度学习模型。

（3）步骤 S3，根据匹配结果确定目标初始化深度学习模型。

根据目标初始化深度学习模型的特征信息在预设的深度学习模型数据库中搜索匹配的结果，确定目标初始化深度学习模型。可以将匹配结果中匹配程度最高的特征信息对应的已训练深度学习模型作为目标初始化深度学习模型。匹配结果中匹配程度最高的已训练深度学习模型，其特征信息与目标初始化深度学习模型的特征信息匹配程度最高，说明二者深度学习模型的各项参数也最相关，直接将其作为该目标初始化深度学习模型，可以在保证初始化结果准确性的前提下，极大缩短目标深度学习模型的初始化时间，优化深度学习模型初始化的结果，从而能够缩短训练时间，提高深度学习模型的训练速度。

（4）步骤 S4，通过训练数据训练目标初始化深度学习模型，得到深度学习的目标深度学习模型。

获得目标初始化深度学习模型后，基于该目标初始化深度学习模型通过训练数据进行训练，得到深度学习的目标深度学习模型。具体步骤为：通过训练数据对搜索匹配确定的目标初始化深度学习模型进行训练时，可以利用训练数据中的输入数据对目标初始化深度学习模型进行无监督训练，再利用训练数据中的输入数据及其对应的输出数据对目标初始化深度学习模型进行有监督训练，从而相当于对已一次训练过的已训练深度学习模型依据目标深度学习模型的需求进行二次训练，得到深度学习的目标深度学习模型。

上述深度学习模型深度学习方法，根据获取的目标初始化深度学习模型的特征信息在预先构建的深度学习模型数据库中进行匹配，根据该匹配的结果确定所需的目标初始化深度学习模型，并通过训练数据对其继续进行训练，最后得到深度学习的目标深度学习模型。本技术方案通过特征信息匹配，将预先存储的以往训练过的深度学习模型作为将来深度学习模型初始化的基础，从而优化深度学习模型初始化的结果，使得深度学习模型在更优初始化的情况下进行训练，从而能够缩短训练时间，提高深度学习模型的训练速度，同时也能保证深度学习模型深度学习的效果。

特征信息可以包括网络标签、数据格式和网络功能中的至少一种。

特征信息用于描述目标初始化深度学习模型。特征信息可以包括一个或多个参数，如可以包括网络标签、数据格式和网络功能中的至少一种。网络标签可以包括是否类标签、真假类标签等；数据格式可以包括图像、文本、音频等格式；网络功能描述深度学习模型的功能应用，如“用于人脸识别”“用于指纹识别”“用于汉语识别”“用于英语识别”等功能应用。

当特征信息包括多个参数时，可以根据目标初始化深度学习模型的实际需求，为特征信息的各参数设定优先级。如当特征信息包括网络标签、数据格式和网络功能时，设定网络功能为高级参数，数据格式为中级参数，网络标签为低级参数，那么当进行特征信息匹配时，优先匹配优先级高的参数，其次匹配优先级低的参数，从而能够有效地加快在深度学习模型数据库中的搜索匹配效率。如特征信息包括：网络标签为“是否类标签”，数据格式为“图片格式”以及网络功能为“人脸识别”，那么在特征信息匹配时，优先匹配网络功能为“人脸识别”的深度学习模型，其次匹配数据格式为“图片格式”的深度学习模型，最后匹配网络标签为“是否类标签”的深度学习模型。

2. 基于大数据的深度学习模型初始化的详细方法

图 2-2 为另一个技术方案中深度学习模型深度学习方法的流程示意图，相比于图 2-1 的技术方案，本技术方案将目标初始化深度学习模型的特征信息在预设深度学习模型数据库中进行匹配，得到匹配结果的步骤包括：

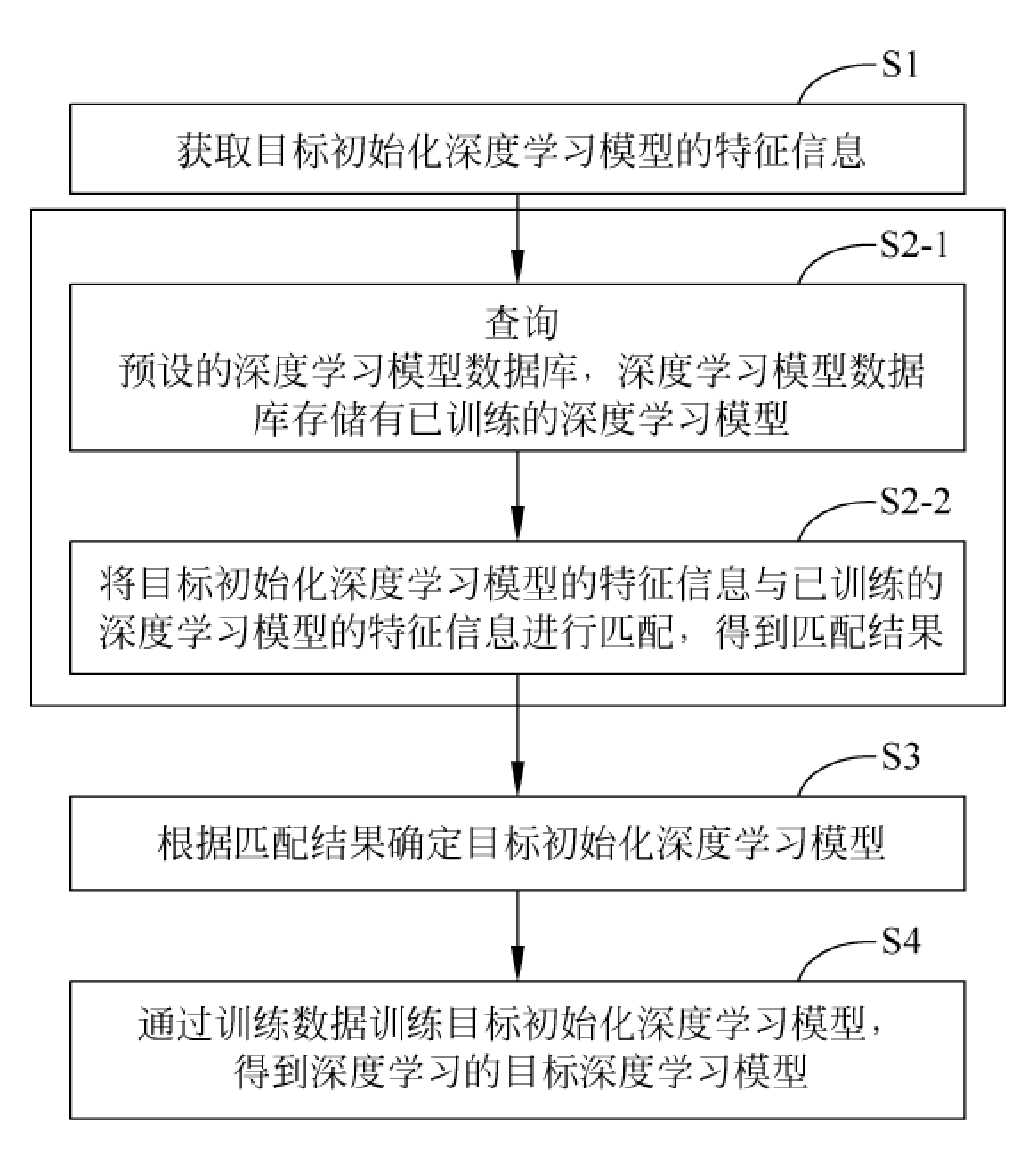


图 2-2　深度学习模型深度学习方法的流程示意图

步骤 S2-1，查询预设的深度学习模型数据库，其中存储有已训练深度学习模型。

深度学习模型数据库包括预先采集的各种已训练深度学习模型的大数据。在获取预设的深度学习模型数据库之前还可以包括：构建深度学习模型数据库，其中存储有已训练深度学习模型。构建深度学习模型数据库可以通过采集各种已训练的深度学习模型来实现，其中，已训练的深度学习模型也设有相应的特征信息。

深度学习模型数据库包括各种已训练深度学习模型，例如，对于人脸识别的深度学习模型，其训练数据的输入/输出格式会有多种，输入格式可以是高分辨率的图像格式、中分辨率的图像格式，也可以是低分辨率的图像格式；输出格式可以是“是/否”标签、“否/是男人”标签，也可以是“否/是女人”标签，还可以是“否/是男人/是女人”标签；应用功能可以是识别人脸、识别男人人脸，还可以是识别女人人脸，还可以是识别男人和女人人脸。以此，根据各

已训练的深度学习模型的特征信息可以细分出各种已训练的子深度学习模型，并构建出一对于人脸识别的深度学习模型子数据库，多种子深度学习模型子数据库进一步组成深度学习模型数据库。

步骤 S2-2，将目标初始化深度学习模型的特征信息与已训练深度学习模型的特征信息进行匹配，得到匹配结果。

获取预设的深度学习模型数据库后，将目标初始化深度学习模型的特征信息与深度学习模型数据库中各已训练深度学习模型的特征信息进行匹配。特征信息进行匹配时，可以一一匹配，也可以通过设计匹配优先级来加快搜索匹配。当特征信息包括多个参数时，可以根据目标初始化深度学习模型的实际需求，为特征信息的各参数设定优先级，当进行特征信息匹配时，优先匹配优先级高的参数，其次匹配优先级低的参数，从而能够有效地加快在深度学习模型数据库中的搜索匹配效率。特征信息进行匹配后，获得各匹配结果。可以将各匹配结果进行排序，以便后续对匹配结果进行处理。

将目标初始化深度学习模型的特征信息与已训练深度学习模型的特征信息进行匹配，得到匹配结果的步骤可以通过以下形式进行：

① 获取已训练深度学习模型的特征信息；

② 获取预设的特征信息匹配规则，其中包括匹配优先级；

③ 根据匹配优先级，将目标初始化深度学习模型的特征信息与已训练深度学习模型的特征信息进行相关度匹配，得到匹配结果。

具体步骤为：首先获取已训练深度学习模型的特征信息，其与目标初始化深度学习模型特征信息的格式相对应，以便实现匹配。获取预设的特征信息匹配规则，其中包括匹配优先级，匹配优先级可以根据目标初始化深度学习模型进行个性化配置，以便在保证匹配准确度的同时，能最大限度地缩短匹配时间，提高匹配效率。特征信息匹配规则还可以包括其他依据目标初始化深度学习模型进行配置的个性化匹配规则，以便快速准确获得匹配度最高的已训练深度学习模型。获得预设的特征信息匹配规则后，根据该匹配规则将目标初始化深度学习模型的特征信息与已训练深度学习模型的特征信息进行相关度匹配。例如，当特征信息包括网络标签、数据格式和网络功能时，首先获取各个已训练深度学习模型的特征信息，再获取预设的特征信息匹配规则，特征信息匹配规则包括的匹配优先级如下：网络功能为高级参数，数据格式为中级参数，网络标签为低级参数。那么，当进行特征信息匹配时，根据该匹配优先级，优先匹配网络功能；当网络功能匹配相同时，匹配数据格式；当数据格式匹配一致时，再匹配网络标签，从而能够有效地加快在深度学习模型数据库中的搜索匹配效率。

例如，一深度学习模型数据库中包括但不限于如下 24 个深度学习模型：

已训练深度学习模型 1，应用信息：训练数据的输入格式是高分辨率 $K11$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别人脸。其中，$K11$ 为分辨率数值。

已训练深度学习模型 2，应用信息：训练数据的输入格式是高分辨率 $K11$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别男人人脸。其中，$K11$ 为分辨率数值。

已训练深度学习模型 3，应用信息：训练数据的输入格式是高分辨率 $K11$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别女人人脸。其中，$K11$ 为分辨率

数值。

已训练深度学习模型 4，应用信息：训练数据的输入格式是高分辨率 $K11$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别张三人脸。其中，$K11$ 为分辨率数值。

已训练深度学习模型 5，应用信息：训练数据的输入格式是高分辨率 $K12$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别人脸。其中，$K12$ 为分辨率数值。

已训练深度学习模型 6，应用信息：训练数据的输入格式是高分辨率 $K12$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别男人人脸。其中，$K12$ 为分辨率数值。

已训练深度学习模型 7，应用信息：训练数据的输入格式是高分辨率 $K13$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别女人人脸。其中，$K13$ 为分辨率数值。

已训练深度学习模型 8，应用信息：训练数据的输入格式是高分辨率 $K14$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别张三人脸。其中，$K14$ 为分辨率数值。

已训练深度学习模型 9，应用信息：训练数据的输入格式是中分辨率 $K21$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别人脸。其中，$K21$ 为分辨率数值。

已训练深度学习模型 10，应用信息：训练数据的输入格式是中分辨率 $K21$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别男人人脸。其中，$K21$ 为分辨率数值。

已训练深度学习模型 11，应用信息：训练数据的输入格式是中分辨率 $K21$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别女人人脸。其中，$K21$ 为分辨率数值。

已训练深度学习模型 12，应用信息：训练数据的输入格式是中分辨率 $K21$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别张三人脸。其中，$K21$ 为分辨率数值。

已训练深度学习模型 13，应用信息：训练数据的输入格式是中分辨率 $K22$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别人脸。其中，$K22$ 为分辨率数值。

已训练深度学习模型 14，应用信息：训练数据的输入格式是中分辨率 $K22$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别男人人脸。其中，$K22$ 为分辨率数值。

已训练深度学习模型 15，应用信息：训练数据的输入格式是中分辨率 $K22$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别女人人脸。其中，$K22$ 为分辨率数值。

已训练深度学习模型 16，应用信息：训练数据的输入格式是中分辨率 $K22$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别张三人脸。其中，$K22$ 为分辨率数值。

已训练深度学习模型 17，应用信息：训练数据的输入格式是低分辨率 $K31$ 的图像格

式；训练数据的输出格式是“是/否”标签；功能可以是识别人脸。其中，$K31$ 为分辨率数值。

已训练深度学习模型 18，应用信息：训练数据的输入格式是低分辨率 $K31$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别男人人脸。其中，$K31$ 为分辨率数值。

已训练深度学习模型 19，应用信息：训练数据的输入格式是低分辨率 $K31$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别女人人脸。其中，$K31$ 为分辨率数值。

已训练深度学习模型 20，应用信息：训练数据的输入格式是低分辨率 $K31$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别张三人脸。其中，$K31$ 为分辨率数值。

已训练深度学习模型 21，应用信息：训练数据的输入格式是低分辨率 $K32$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别人脸。其中，$K32$ 为分辨率数值。

已训练深度学习模型 22，应用信息：训练数据的输入格式是低分辨率 $K32$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别男人人脸。其中，$K32$ 为分辨率数值。

已训练深度学习模型 23，应用信息：训练数据的输入格式是低分辨率 $K32$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别女人人脸。其中，$K32$ 为分辨率数值。

已训练深度学习模型 24，应用信息：训练数据的输入格式是低分辨率 $K32$ 的图像格式；训练数据的输出格式是“是/否”标签；功能可以是识别张三人脸。其中，$K32$ 为分辨率数值。

在匹配时，将目标初始化深度学习模型的特征信息与深度学习模型数据库中已训练深度学习模型 1～24 进行匹配。具体而言，可以匹配训练数据的输入格式、训练数据的输出格式及网络功能等。

根据匹配结果确定目标初始化深度学习模型可以通过以下步骤进行：

① 将相关度匹配的匹配结果进行排序；

② 获取相关度匹配的匹配结果中相关度最高的已训练深度学习模型的特征信息所对应的已训练深度学习模型；

③ 将相关度最高的已训练深度学习模型作为目标初始化深度学习模型。

其中，将目标初始化深度学习模型的特征信息与已训练深度学习模型的特征信息进行相关度匹配的匹配结果进行排序，如按照相关度由高至低的顺序进行排序，根据该匹配结果中相关度最高的特征信息获取其对应的已训练深度学习模型，最后将该相关度最高的已训练深度学习模型作为目标初始化深度学习模型。

例如，对于上述深度学习模型数据库，其至少包括上述列出的已训练深度学习模型 1～24。如果此时需要待初始化的目标初始化深度学习模型的特征信息是：训练数据的输入输出格式为“图片格式且分辨率为 $K12$、是否类标签”，功能为“人脸识别”，则与从上述深度学习模型数据库中搜索出来的特征信息最匹配的已训练深度学习模型是已训练深度学习模型 5，

将该模型作为所需的目标初始化深度学习模型。

如果需要待初始化的目标初始化深度学习模型的特征信息是：训练数据的输入/输出格式为“图片格式且分辨率为 $K32$、是否类标签”，功能为“张三人脸识别”，则与从上述深度学习模型数据库中搜索出来的特征信息最匹配的已训练深度学习模型是已训练深度学习模型 24，将该模型作为所需的目标初始化深度学习模型。

当未搜索匹配到完全相同对应的已训练深度学习模型时，可以从匹配结果中提取相似度最高的特征信息对应的已训练深度学习模型，将其作为所需的目标初始化深度学习模型。

例如，如果需要待初始化的目标初始化深度学习模型的特征信息是：训练数据的输入/输出格式为“图片格式且分辨率为 $X1$、是否类标签”，功能为“人脸识别”，其中，$X1$ 与 $K12$ 最为接近，则与从上述深度学习模型数据库中搜索出来的特征信息最匹配的已训练深度学习模型是已训练深度学习模型 5，那么将该模型作为所需的目标初始化深度学习模型。

如果需要待初始化的目标初始化深度学习模型的特征信息是：训练数据的输入/输出格式为“图片格式且分辨率为 $X2$、是否类标签”，功能为“张三人脸识别”，其中，$X2$ 与 $K32$ 最为接近，则与从上述深度学习模型数据库中搜索出来的特征信息最匹配的已训练深度学习模型是已训练深度学习模型 24，那么将该模型作为所需的目标初始化深度学习模型。

在根据匹配结果确定目标初始化深度学习模型后，通过训练数据训练该目标初始化深度学习模型，得到深度学习的目标深度学习模型，该模型可以用于具体的应用中。例如，如果需要待初始化的目标初始化深度学习模型的特征信息是：训练数据的输入/输出格式为“图片格式且分辨率为 $X1$、是否类标签”，功能为“人脸识别”，其中，$X1$ 与 $K12$ 最为接近，则与从上述深度学习模型数据库中搜索出来的特征信息最匹配的已训练深度学习模型是已训练深度学习模型 5，那么将该模型作为目标初始化深度学习模型，对其继续进行训练。将输入/输出格式为“图片格式且分辨率为 $X1$、是否类标签”的训练数据输入已训练深度学习模型 5 进行训练，得到新的已训练深度学习模型 A，则可应用已训练深度学习模型 A 进行“人脸识别”测试和应用。

再如，如果需要待初始化的目标初始化深度学习模型的特征信息是：训练数据的输入/输出格式为“图片格式且分辨率为 $X2$、是否类标签”，功能为“张三人脸识别”，其中，$X2$ 与 $K32$ 最为接近，则与从上述深度学习模型数据库中搜索出来的特征信息最匹配的已训练深度学习模型是已训练深度学习模型 24，那么将该模型作为目标初始化深度学习模型，对其继续进行训练。将输入/输出格式为“图片格式且分辨率为 $X2$、是否类标签”的训练数据输入已训练深度学习模型 24 进行训练，得到新的已训练深度学习模型 B，则可应用已训练深度学习模型 B 进行“张三人脸识别”测试和应用。

本技术方案的深度学习模型深度学习方法包括：将深度学习的目标深度学习模型更新至深度学习模型数据库中。

在对目标初始化深度学习模型训练结束，得到深度学习的目标深度学习模型后，将其更新添加至深度学习模型数据库中，以及时扩展深度学习模型数据库，以便后续深度学习模型搜索匹配使用。

例如，得到新的已训练深度学习模型 A 后，则将其 A 及其特征信息添加进深度学习模型数据库中，已训练深度学习模型 A 的特征信息是：训练数据的输入/输出格式为“图片格式且分辨率为 $X1$、是否类标签”，功能为“人脸识别”。

又如，得到新的已训练深度学习模型B后，则将其及其特征信息添加进深度学习模型数据库，已训练深度学习模型B的特征信息是：训练数据的输入/输出格式为“图片格式且分辨率为 $X2$、是否类标签”，功能为“张三人脸识别”。

如此，后续目标初始化深度学习模型的特征信息如果为“训练数据的输入/输出格式为‘图片格式且分辨率为 $X1$、是否类标签’，功能为‘人脸识别’”或“训练数据的输入/输出格式为‘图片格式且分辨率为 $X2$、是否类标签’，功能为‘张三人脸识别’”时，可以快速准确地匹配得到已训练深度学习模型A或B，并作为新的目标初始化深度学习模型，而不需要再重新以随机赋予初值的方式进行初始化，能够优化深度学习模型初始化的结果，从而缩短训练时间，提高深度学习模型的训练速度，同时也能保证深度学习模型深度学习的效果。

2.1.2 基于大数据的深度学习模型初始化系统

基于上述深度学习模型深度学习方法，本技术方案还提供一种深度学习模型深度学习系统。

深度学习模型深度学习系统的结构框图如图2-3所示，本实现中的深度学习模型深度学习系统包括：

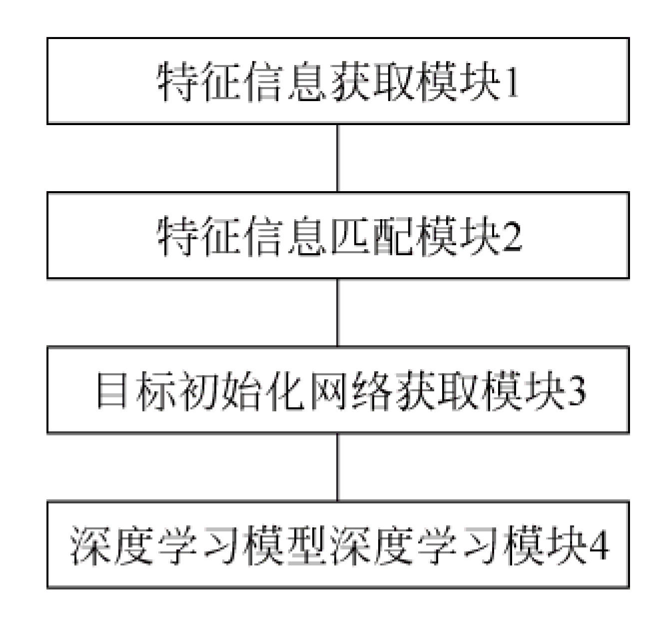


图2-3 深度学习模型深度学习系统的结构框图

(1) 特征信息获取模块1，用于获取目标初始化深度学习模型的特征信息。

(2) 特征信息匹配模块2，用于将目标初始化深度学习模型的特征信息在预设深度学习模型数据库中进行匹配，得到匹配结果。

特征信息匹配模块2可以包括：

① 数据库获取单元2-1，用于查询预设的深度学习模型数据库，深度学习模型数据库存储有已训练深度学习模型；

② 匹配单元2-2，用于将目标初始化深度学习模型的特征信息与已训练深度学习模型的特征信息进行匹配，得到匹配结果。

(3) 目标初始化网络获取模块3，用于根据匹配结果确定目标初始化深度学习模型；

(4) 深度学习模型深度学习模块4，用于通过训练数据训练目标初始化深度学习模型，得到深度学习的目标深度学习模型。

上述深度学习模型深度学习系统，由特征信息获取模块获取目标初始化深度学习模型的特征信息，由特征信息匹配模块根据获取的目标初始化深度学习模型的特征信息在预先构建的深度学习模型数据库中进行匹配，得到匹配结果，再由目标初始化网络获取模块根据该匹配结果确定所需的目标初始化深度学习模型，最后由深度学习模型深度学习模块通过训练数据对该目标初始化深度学习模型继续进行训练，得到深度学习的目标深度学习模型。本技术方案的深度学习模型深度学习系统通过特征信息匹配，将预先存储的以往训练过的深度学习模型作为将来深度学习模型初始化的基础，从而优化深度学习模型初始化的结果，使得深度学习模型在更优初始化的情况下进行训练，从而能够缩短训练时间，提高深度学习模型的训练速度，同时也能保证深度学习模型深度学习的效果。

2.2　深度学习模型进化

神经网络是一种进行分布式并行信息处理的算法数学模型，常用于智能机器识别。深度学习模型是通过样本数据训练进行深度学习并通过测试之后得到的神经网络，其中，深度学习分有监督学习和无监督学习。

通常在得到深度学习模型之后即不会更改，也就是在已训练深度学习模型的使用过程中不会改变。然而，深度学习模型的训练和测试的数据毕竟有限，准确率不能保证，且随着已训练深度学习模型使用过程中输入的数据随时间发生变化，当使用时间较长时，深度学习模型的准确率降低。例如，识别某个人的人脸识别深度学习模型，当这个人随着年龄增长，其脸部特征也在不断发生变化，所以该人脸识别深度学习模型的适用性在不断变差，准确率降低[9,10]。

本技术研究并给出了一种新的深度学习模型进化方法、装置、介质和计算机设备。该方法包括：将输入数据输入调取的已训练的深度学习模型得到输出数据，记录使用时间及获取对应的真实结果数据；按照预设的抽查规则选取使用过程中的输入数据、输出数据、使用时间和真实结果数据；将选取的输入数据及对应的输出数据、真实结果数据和使用时间组成数据组；计算数据组中输出数据与真实结果数据的差异值，统计预设时间段内所有数据组中差异值大于预设阈值的差异组的组数，若组数大于或等于存储的组数阈值，根据差异组的输入数据和真实结果数据对已训练的深度学习模型进行训练得到进化后的深度学习模型。如此可纠正深度学习模型的偏差，提高准确率。

2.2.1　深度学习模型进化方法

1. 深度学习模型进化的基本方法

参考图 2-4，在一个技术方案中，提供了一种深度学习模型进化方法，包括如下步骤：

（1）步骤 S1，指调取已训练的深度学习模型，将输入数据输入已训练的深度学习模型进行深度学习，得到输出数据并记录使用时间及获取对应输入数据的真实结果数据。

输入数据是用于输入深度学习模型的数据，在步骤 S110 之前包括获取输入数据步骤。输入数据可以通过采集或用户输入得到，输入数据的类型由对应的深度学习模型的类型确定。例如，若深度学习模型为用于人脸识别的神经网络，则输入数据为人脸图像数据；若深度学习模型为用于语音识别的神经网络，则输入数据为语音数据。

其中，使用时间是将输入数据输入已训练的深度学习模型的时间，也可以是得到输出数据的时间。使用时间可以以日期为单位，也可以精确到日期内的具体时刻。对应的真实结果数据是对应输入数据的客观现实结果，可以通过用户反馈得到；对应的真实结果数据也可以根据人工检测输入数据得到。例如，输入数据为人脸图像数据，则可以直接由用户反馈“0”或“1”作为对应的真实结果数据，也可以通过人工检测得到“0”或“1”，“0”和“1”分别表示输入的人脸图像数据对应为非目标人物的人脸、目标人物的人脸。

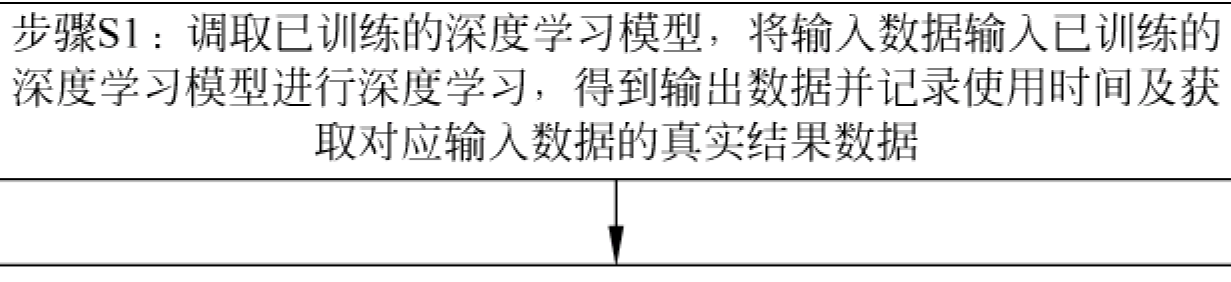


图 2-4　深度学习模型进化方法的流程

已训练的深度学习模型是需要进行进化的深度学习模型。其中，输出数据与对应的真实结果数据的格式相同，便于比较。例如，输出数据和真实结果数据均用“0”和“1”分别表示输入的人脸图像数据对应为非目标人物的人脸、目标人物的人脸。

（2）步骤 S2，按照预设的抽查规则选取已训练的深度学习模型使用过程中的输入数据、输出数据、使用时间和真实结果数据。

步骤 S1 中将输入数据输入已训练的深度学习模型是对已训练的深度学习模型的使用，多次进行步骤 S1 则可以得到每一次使用对应的输入数据、输出数据、使用时间和真实结果数据。如果每次使用已训练的深度学习模型所对应的数据都获取，会导致数据处理成本过高，所以需要采用抽查的方式来降低数据处理成本。例如，预设的抽查规则可以是每 K 次使用中获取 1 次使用的输入数据、输出数据、使用时间和对应的真实结果数据，其中，$K>1$，例如 100。

（3）步骤 S3，将选取的输入数据及对应的输出数据、对应的真实结果数据和对应的使用时间组成一个数据组。

例如，S3 可以包括：将 S2 中每次获取到的输入数据、输出数据、真实结果数据和使用时间组成一个数据组并保存。可以将数据组保存在数据库中，如存入进化知识库的一个数据组：1 月 1 日，输入数据为人脸图像 A，输出数据为 1（代表张三人脸），对应的真实结果数据为 1（代表张三人脸）。

（4）步骤 S4，计算每个数据组中输出数据与同组内对应的真实结果数据的差异值，以使用时间属于预设时间段内的所有数据组中对应差异值大于预设阈值的数据组为差异组，统计差异组的组数。

获取的同组内的输出数据与对应的真实结果数据的差异值可以反映已训练的深度学习模型此次深度学习的准确性。比较输出数据与对应的真实结果数据是否一致，越一致，则差

异值越小;越不一致,则差异值越大。预设时间段和预设阈值根据实际需要预先设置。预设阈值可以为0,也可以为能够接受的输出误差值。差异值大于预设阈值,表示对应数据组中的输出数据和对应的真实结果数据的差异不可接受,即输出结果错误,从而统计的差异组的组数即为预设时间段内输出结果错误的次数。

(5) 步骤S5,若组数大于或等于存储的组数阈值,则将每一个差异组中的输入数据和真实结果数据分别作为输入和预期输出对已训练的深度学习模型进行训练,得到进化后的深度学习模型。

组数阈值用于作为参考,若组数大于或等于组数阈值,则表示预设时间段内深度学习结果错误的次数不在可接受范围内,已训练的深度学习模型的准确率较低,此时通过将每一个差异组的输入数据和真实结果数据分别作为输入和预期输出对已训练的深度学习模型进行训练,可进化深度学习模型。

上述深度学习模型进化方法,通过将输入数据输入已训练的深度学习模型进行深度学习得到输出数据,选取输入数据及与输入数据对应的输出数据、使用时间和真实结果数据并将相对应的数据作为一个数据组,统计使用时间属于预设时间段内的所有数据组中输出数据与对应的真实结果数据的差异值大于预设阈值的差异组的组数;在组数大于或等于组数阈值时,采用差异组的输入数据和对应的真实结果数据对已训练的深度学习网络进行训练,从而在已训练的深度学习模型的使用过程中进行进化,可纠正深度学习模型的偏差,提高深度学习模型的准确率。

2. 深度学习模型进化的详细方法

在一个技术方案中,步骤S4之后包括:若组数小于存储的组数阈值,则返回步骤S1。

差异组的组数小于组数阈值,表示预设时间段内存在差异的数据组的数量在可接受范围内,此时返回步骤S1,可继续使用已训练的深度学习模型。

具体而言,每按照预设的抽查规则完成一次选取可保存一个数据组。在一个技术方案中,步骤S4可以包括:计算最新保存的数据组中输出数据与对应的真实结果数据的差异值,若差异值大于预设阈值,则更新已存的差异组的组数。其中,已存的差异组的组数初始值为0,更新已存的差异组的组数可以是在已存的差异组的组数上加1得到新的差异组的组数。

通过在按照预设的抽查规则完成每一次选取后统计差异组的组数,后续深度学习完成后只需要在已存的差异组的组数上简单地更新即可,处理简便。

也可以采用其他方式统计差异组的组数,例如,步骤S4还可以包括:获取所有数据组中输出数据与同组内对应的真实结果数据的差异值,以使用时间属于预设时间段内的所有数据组中对应差异值大于预设阈值的数据组为差异组,统计差异组的组数。即:每按照预设的抽查规则完成一次选取,则统计使用时间属于预设时间段内的所有数据组中差异组的组数。

在一个技术方案中,组数阈值可以为固定值。将差异组的组数与固定值比较以纠正深度学习模型的偏差,固定值越小,可接受的错误次数越少,对应进化的深度学习模型准确性越高。通过预设一个固定值并存储作为组数阈值,固定值可以根据实际需要具体设置,因此可提高进化效果设计的灵活性。

在另一个技术方案中,步骤S3之后、步骤S5之前,包括组数阈值存储步骤,即根据数据

组获取当前的组数阈值并存储。

根据得到的数据组获取组数阈值，组数阈值随存储的数据组动态调整，适应性高。具体而言，组数阈值存储步骤可以是在步骤 S4 之前执行，也可以是在步骤 S4 之后执行，还可以与步骤 S4 同时执行。

组数阈值存储步骤可以包括：统计使用时间属于预设时间段内的数据组的总组数；根据总组数计算得到当前的组数阈值并存储。其中，当前的组数阈值小于总组数。

根据总组数计算得到当前的组数阈值，得到的组数阈值随总组数变动，即随预设时间段内深度学习的总次数变动，从而可根据总组数对应的组数阈值判断是否需要对已训练的深度学习模型进行训练。

根据总组数计算得到当前的组数阈值并存储可以是计算总组数与第一预设比例的乘积得到当前的组数阈值并存储，其中，第一预设比例大于零且小于 1。本技术方案中，第一预设比例为 1/4。

组数阈值存储步骤还可以包括：统计使用时间属于预设参考时间段内的所有数据组中对应差异值大于预设阈值的数据组的数量；根据数据组数量计算得到当前的组数阈值并存储。

预设参考时间段可以根据实际需要设置。根据统计的数量计算得到当前的组数阈值，组数阈值随统计得到的数量变动；如果将统计得到的数量直接作为当前的组数阈值，则相当于将预设参考时间段内与预设时间段内的抽查得到的输出结果的错误次数进行比较。当预设时间段内抽查到的输出结果的错误次数大于或等于预设参考时间段内抽查到的输出结果的错误次数时，采用差异组的输入数据和真实结果数据对已训练的深度学习模型进行训练。

可以计算统计得到的数量与第二预设比例的乘积得到当前的组数阈值并存储。当第二预设比例大于 1 时，例如为 1.5，表明要求深度学习模型的准确率在使用过程中需要变高到相应比例。当第二预设比例小于 1 时，例如为 0.5，表明对深度学习模型的准确率要求在使用的过程中可以变低到相应比例。

在一个技术方案中，预设时间段为以当前时刻为结束点的对已训练的深度学习模型进行预设次数的使用的时间段；或者为以当前时刻为结束点、时长为第一预设时长的时间段，即时长固定。例如，预设时间段可以是从开始到当前时刻的整个时间段，可以是整个时间段中最近 1 年的时间段，也可以是整个时间段中最近 100 次使用已训练的深度学习模型的时间段。其中，预设次数和第一预设时长可以根据实际情况具体设置。

本技术方案中，预设参考时间段为以预设起始时刻为起始点、预设结束时刻为结束点的时间段，即预设参考时间段是预先设置的一个固定时间段；或者可以是一个以已训练的深度学习模型第一次使用的时刻为起始点、时长为第一预设时长的时间段；或者可以是一个以已训练的深度学习模型第一次使用的时刻为起始点、使用次数为预设次数的时间段；或者可以是一个以当前时刻的第二预设时长之前的时刻为结束点、时长为第一预设时长的时间段；或者可以是一个以当前时刻的第二预设时长之前的时刻为结束点、使用次数为预设次数的时间段。其中，第二预设时长大于第一预设时长。

在一个技术方案中，步骤 S5 中将每一个差异组中的输入数据和真实结果数据分别作为输入和预期输出对已训练的深度学习模型进行训练，得到进化后的深度学习模型的步骤

包括：

① 步骤 5-1，将所有差异组中的输入数据输入已训练的深度学习模型进行无监督学习，得到初步进化的深度学习模型。

② 步骤 5-2，将所有差异组中的输入数据和对应的真实结果数据分别作为输入和预期输出，对初步进化的深度学习模型进行有监督学习，得到进化后的深度学习模型。

无监督学习和有监督学习是深度学习的学习方式。在将所有差异组中的输入数据输入已训练的深度学习模型进行无监督学习得到初步进化的深度学习模型的基础上，再根据差异组中的输入数据和对应的真实结果数据对初步进化的深度学习模型进行有监督学习，结合无监督学习和有监督学习，得到的进化后的深度学习模型更接近全局最优，偏差纠正效果好，从而准确率更高。

在一个技术方案中，参考图 2-5，步骤 S5-1 包括步骤 S5-1-1 和步骤 S5-1-2。

步骤S5-1-1：按照各差异组使用时间的先后顺序排序，选取第一差异组中的输入数据输入已训练的深度学习模型进行无监督学习，得到对应的训练深度学习模型

↓

步骤S5-1-2：选取下一差异组中的输入数据作为前一差异组对应的训练深度学习模型的输入，对前一差异组对应的训练深度学习模型进行无监督学习得到对应的训练深度学习模型，循环直到得到最后一组差异组对应的训练深度学习模型并作为初步进化的深度学习模型

↓

步骤S5-2-1：按照各差异组使用时间的先后顺序排序，选取第一差异组中的输入数据和对应的真实结果数据分别作为初步进化的深度学习模型的输入和预期输出，对初步进化的深度学习模型进行有监督学习，得到对应的改进深度学习模型

↓

步骤S5-2-2：选取下一差异组中的输入数据和真实结果数据分别作为前一差异组对应的改进深度学习模型的输入和预期输出，对前一差异组对应的改进深度学习模型进行有监督学习，得到改进深度学习模型，循环直到得到最后一组差异组对应的改进深度学习模型并作为进化后的深度学习模型

图 2-5　得到进化后的深度学习模型的具体流程

步骤 S5-1-1，按照各差异组使用时间的先后顺序排序，选取第一差异组中的输入数据输入已训练的深度学习模型进行无监督学习，得到对应的训练深度学习模型。

差异组的组数有多个时，按照差异组使用时间的先后顺序进行排序，使用时间越靠前的排序越靠前，第一差异组为最先进行抽查得到的差异组。

步骤 S5-1-2，选取下一差异组中的输入数据作为前一差异组对应的训练深度学习模型的输入，对前一差异组对应的训练深度学习模型进行无监督学习得到对应的训练深度学习模型，循环直到得到最后一组差异组对应的训练深度学习模型并作为初步进化的深度学习模型。

按照使用时间的先后顺序进行无监督学习，后一次无监督学习在前一次无监督学习得到的训练深度学习模型上进行，从而使用时间越靠后越接近将来使用的时间，即越是后输入的数据越接近将来的使用时间，而越后输入的数据是过去最近的时间产生的数据，从而使得无监督学习后的深度学习模型更能适应将来使用的需要。

在一个技术方案中,继续参考图 2-5,步骤 S5-2 包括步骤 S5-2-1 和步骤 S5-2-2。

步骤 S5-2-1,按照各差异组使用时间的先后顺序排序,选取第一差异组中的输入数据和对应的真实结果数据分别作为初步进化的深度学习模型的输入和预期输出,对初步进化的深度学习模型进行有监督学习,得到对应的改进深度学习模型。

步骤 S5-2-2,选取下一差异组中的输入数据和真实结果数据分别作为前一差异组对应的改进深度学习模型的输入和预期输出,对前一差异组对应的改进深度学习模型进行有监督学习,得到改进深度学习模型,循环直到得到最后一组差异组对应的改进深度学习模型并作为进化后的深度学习模型。

在无监督学习后的初步进化的深度学习模型基础上,按照使用时间的先后顺序进行有监督学习,后一次有监督学习在前一次有监督学习得到的改进深度学习模型上进行,从而使用时间越靠后越接近将来使用的时间,即越是后输入的数据越接近将来的使用时间,而越后输入的数据是过去最近的时间产生的数据,从而使得最终进化后的深度学习模型更能适应将来使用的需要,准确率更高。

参考图 2-6 和图 2-7,把学习结构看作一个网络,深度学习的核心思路如下:

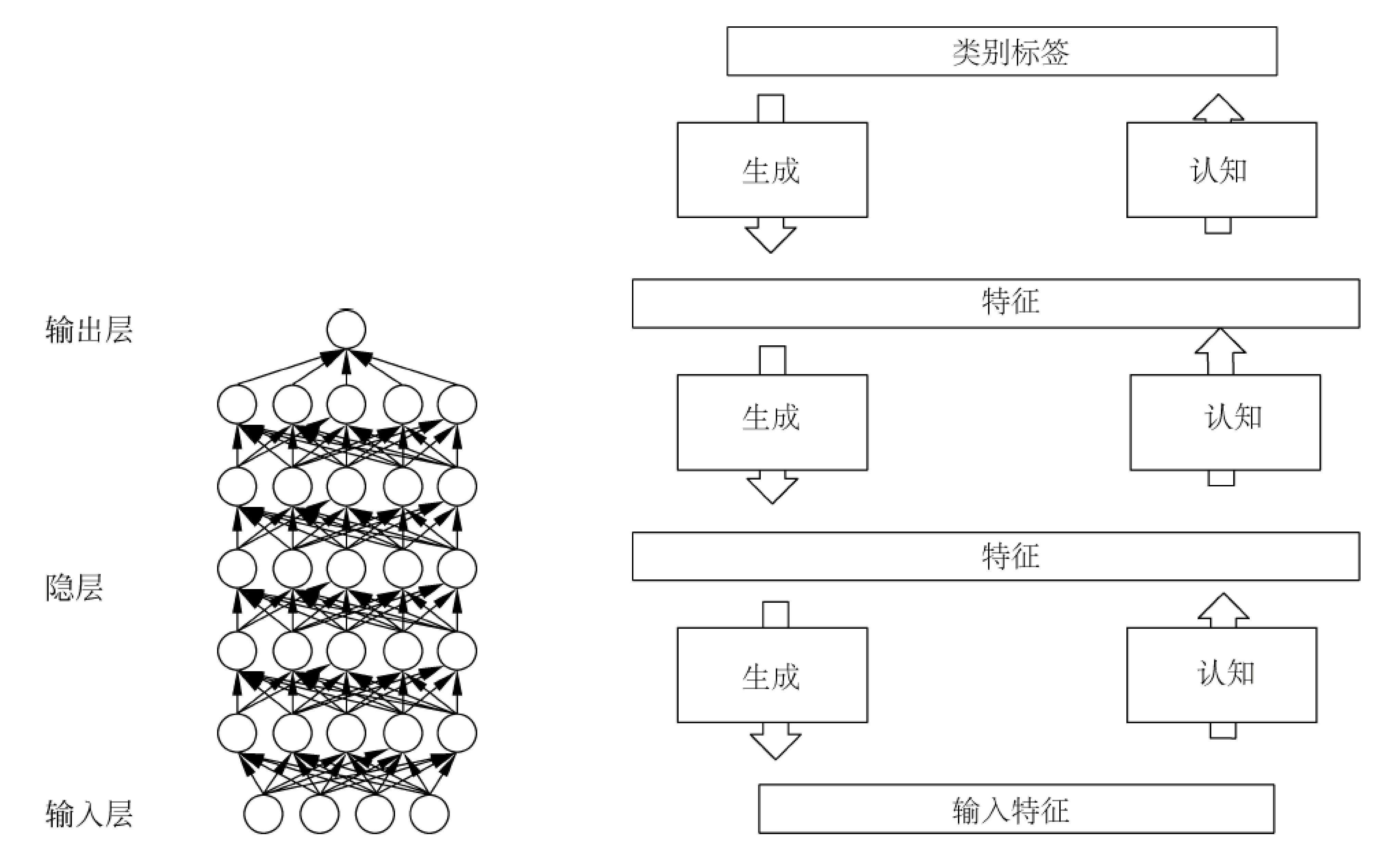


图 2-6　含多个隐藏层的深度学习模型　　　图 2-7　深度学习模型的核心思路示意

首先,采用自下而上的无监督学习,逐层构建单层神经元。每层采用 Wake-Sleep 算法进行调优,每次仅调整一层,逐层调整。这个过程可以看作是一个 Feature Learning 的过程,是和传统神经网络区别最大的部分。其中,Wake-Sleep 算法包括 Wake 阶段和 Sleep 阶段。Wake 阶段:认知过程,通过下层的输入特征(Input)和向上的认知(Encoder)权重产生每一层的抽象表示(Code),再通过当前的生成(Decoder)权重产生一个重建信息(Reconstruction),计算输入特征和重建信息残差,使用梯度下降修改层间的下行生成权重。即 Sleep 阶段:生成过程,通过上层概念和向下的生成权重,生成下层的状态,再利用认知权重产生一个抽象景象。利用初始上层概念和新建抽象景象的残差,利用梯度下降修改层间

向上的认知权重。具体而言,先用无标定数据训练第一层,训练时先学习第一层的参数(这一层可以看作是得到一个使得输出数据和输入数据差别最小的三层神经网络的隐藏层),由于模型 Capacity 的限制以及稀疏性约束,使得得到的模型能够学习到数据本身的结构,从而得到比输入数据更具有表示能力的特征;在学习得到第 $n-1$ 层后,将第 $n-1$ 层的输出数据作为第 n 层的输入数据,训练第 n 层,由此分别得到各层的参数。

然后,进行自顶向下的有监督学习。这一步是在第一步学习获得各层参数进的基础上,在最顶部的编码层添加一个分类器(例如 Logistic 回归、SVM 等),而后通过带标签数据的有监督学习,利用梯度下降法微调整个网络参数。深度学习的第一步实质上是一个网络参数初始化过程。区别于传统神经网络初值随机初始化,深度学习模型是通过在无监督学习输入数据的结构上得到的,因而这个初值更接近全局最优,从而能够取得更好的效果。

在一个技术方案中,步骤 S5 之后包括:删除所有数据组,将进化后的深度学习模型作为新的已训练的深度学习模型,并返回步骤 S1。

通过在步骤 S5 后删除所有数据组,从而进行清零,避免存储的数据组对深度学习模型的下一轮进化产生影响。

2.2.2　深度学习模型进化系统

1. 深度学习模型进化系统的基本模块

参考图 2-8,在一个技术方案中,提供一种深度学习模型进化装置,包括数据获取模块 1、数据选取模块 2、数据保存模块 3、组数统计模块 4 和网络进化模块 5。

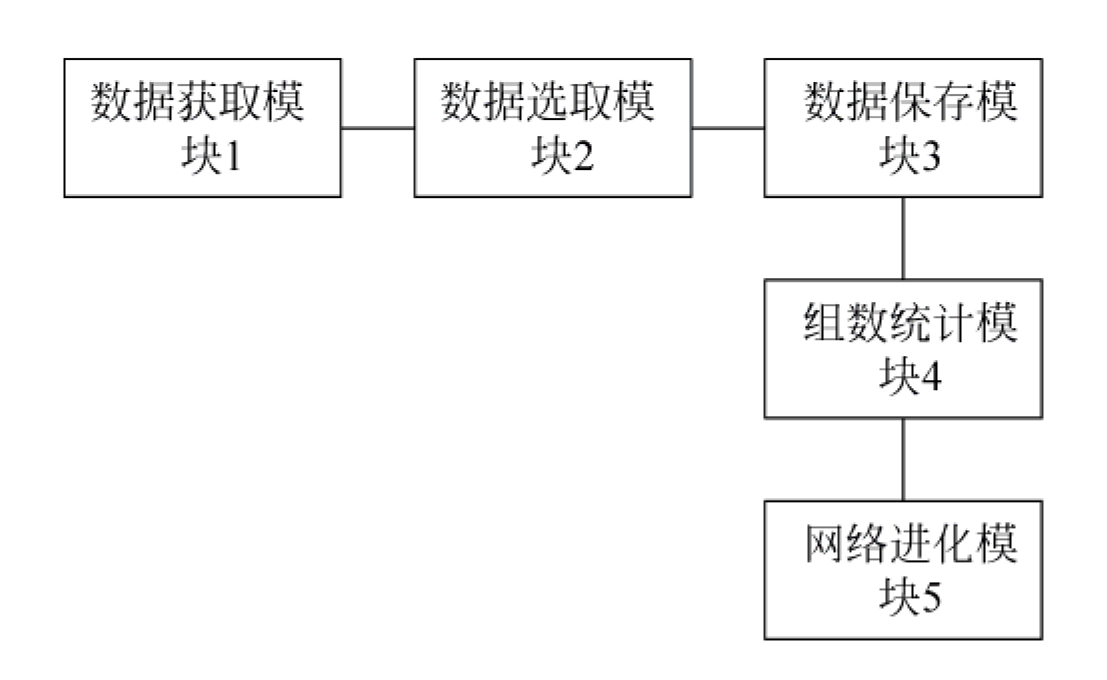


图 2-8　深度学习模型进化装置的结构

(1) 数据获取模块 1,用于调取已训练的深度学习模型,将输入数据输入已训练的深度学习模型进行深度学习,得到输出数据并记录使用时间及获取对应输入数据的真实结果数据。

(2) 数据选取模块 2,用于按照预设的抽查规则选取已训练的深度学习模型使用过程中的输入数据、输出数据、使用时间和真实结果数据。

(3) 数据保存模块 3,用于将选取的输入数据及对应的输出数据、真实结果数据和使用时间组成一个数据组。

(4) 组数统计模块 4,用于计算每个数据组中输出数据与同组内对应的真实结果数据的

差异值,以使用时间属于预设时间段内的所有数据组中对应差异值大于预设阈值的数据组为差异组,统计差异组的组数。

(5) 网络进化模块 5,用于在组数大于或等于存储的组数阈值时,将每一个差异组中的输入数据和真实结果数据分别作为输入和预期输出对已训练的深度学习模型进行训练,得到进化后的深度学习模型。

上述深度学习模型进化装置,通过将输入数据输入已训练的深度学习模型进行深度学习得到输出数据,选取输入数据及与输入数据对应的输出数据、使用时间和真实结果数据并将相对应的数据作为一个数据组,统计使用时间属于预设时间段内的所有数据组中输出数据与对应的真实结果数据的差异值大于预设阈值的差异组的组数;在组数大于或等于组数阈值时,采用差异组的输入数据和对应的真实结果数据对已训练的深度学习网络进行训练,从而在已训练的深度学习模型使用过程中进行进化,可纠正深度学习模型的偏差,提高深度学习模型的准确率。

2. 深度学习模型进化系统的扩展模块

上述深度学习模型进化装置包括循环模块(图 2-8 中未说明),用于在差异组的组数小于存储的组数阈值时,返回数据获取模块 1 再次调取已训练的深度学习模型,将输入数据输入已训练的深度学习模型进行深度学习,得到输出数据并记录使用时间及获取对应输入数据的真实结果数据。如此,可进行下一次使用。

具体而言,每按照预设的抽查规则完成一次选取可保存一个数据组。在一个技术方案中,组数统计模块 4 可以用于计算最近保存的数据组中输出数据与对应的真实结果数据的差异值,若差异值大于预设阈值,则更新已存的差异组的组数。其中,已存的差异组的组数初始值为 0,更新已存的差异组的组数可以是在已存的差异组的组数上加 1 得到新的差异组的组数。

每按照预设的抽查规则完成一次选取后统计差异组的组数,后续深度学习完成后只需要在已存的差异组的组数上简单地更新即可,处理简便。

也可以采用其他方式统计差异组的组数,例如,组数统计模块 4 还可用于:获取所有数据组中输出数据与同组内对应的真实结果数据的差异值,以使用时间属于预设时间段内的所有数据组中对应差异值大于预设阈值的数据组为差异组,统计差异组的组数。即:每按照预设的抽查规则完成一次巡区,则统计所有数据组中差异组的组数。

在一个技术方案中,组数阈值可以为固定值。通过预设一个固定值并存储作为组数阈值,固定值可以根据实际需要具体设置,因此可提高进化效果设计的灵活性。

在另一个技术方案中,上述深度学习模型进化装置包括组数阈值存储模块(图 2-8 中未说明),用于在数据保存模块 3 执行相应功能之后、网络进化模块 5 执行相应功能之前,根据数据组获取当前的组数阈值并存储。根据已经得到的数据组获取组数阈值,组数阈值随存储的数据组动态调整,适应性高。

组数阈值存储模块可以用于:统计使用时间属于预设时间段内的数据组的总组数;根据总组数计算得到当前的组数阈值并存储。其中,当前的组数阈值小于总组数。

根据总组数计算得到当前的组数阈值,组数阈值随总组数变动,即随预设时间段内深度学习的总次数变动,从而可根据总组数对应的组数阈值判断是否需要对已训练的深度学习模型进行训练。

可以计算总组数与第一预设比例的乘积得到当前的组数阈值并存储，其中，第一预设比例大于零且小于1。本技术方案中，第一预设比例为1/4。

组数阈值存储模块还可以用于：统计使用时间属于预设参考时间段内的所有数据组中对应差异值大于预设阈值的数据组的数量；根据数据组数量计算得到当前的组数阈值并存储。

预设参考时间段可以根据实际需要设置，可以是预先设置的一个确定起始时刻、结束时刻的固定时间段，也可以是一个以当前时刻为结束点确定的时间段。通过根据统计的数量计算得到当前的组数阈值，组数阈值随数量变动，从而将两个时间段内的深度学习的错误次数进行比较，当预设时间段内深度学习的错误次数大于或等于预设参考时间段内深度学习的错误次数时，采用差异组的输入数据和真实结果数据对已训练的深度学习模型进行训练。

具体来说，可以由计算数量与第二预设比例的乘积得到当前的组数阈值并存储。其中，当前的组数阈值大于或等于数量。本技术方案中，第二预设比例为1.5。

在一个技术方案中，网络进化模块5将所有差异组中的输入数据输入已训练的深度学习模型进行无监督学习，得到初步进化的深度学习模型；将所有差异组中的输入数据和对应的真实结果数据分别作为输入和预期输出，对初步进化的深度学习模型进行有监督学习，得到进化后的深度学习模型。

在将差异组中的输入数据输入已训练的深度学习模型进行无监督学习得到初步进化的深度学习模型的基础上，再根据差异组中的输入数据和对应的真实结果数据对初步进化的深度学习模型进行有监督学习，结合无监督学习和有监督学习，得到的进化后的深度学习模型更接近全局最优，偏差纠正效果好，从而准确率更高。

在一个技术方案中，网络进化模块5按照各差异组使用时间的先后顺序排序，选取第一差异组中的输入数据输入已训练的深度学习模型进行无监督学习，得到对应的训练深度学习模型；选取下一差异组中的输入数据作为前一差异组对应的训练深度学习模型的输入，对前一差异组对应的训练深度学习模型进行无监督学习得到对应的训练深度学习模型，循环直到得到最后一组差异组对应的训练深度学习模型并作为初步进化的深度学习模型。

按照使用时间的先后顺序进行无监督学习，后一次无监督学习在前一次无监督学习得到的训练深度学习模型上进行，从而使用时间越靠后越接近将来使用的时间，即越是后输入的数据越接近将来的使用时间，而越后输入的数据是过去最近的时间产生的数据，从而使得无监督学习后的深度学习模型更能适应将来使用的需要。

在一个技术方案中，网络进化模块5按照各差异组使用时间的先后顺序排序，选取第一差异组中的输入数据和对应的真实结果数据分别作为初步进化的深度学习模型的输入和预期输出，对初步进化的深度学习模型进行有监督学习，得到对应的改进深度学习模型；选取下一差异组中的输入数据和真实结果数据分别作为前一差异组对应的改进深度学习模型的输入和预期输出，对前一差异组对应的改进深度学习模型进行有监督学习，得到改进深度学习模型，循环直到得到最后一组差异组对应的改进深度学习模型并作为进化后的深度学习模型。

在无监督学习后的初步进化的深度学习模型的基础上，按照使用时间的先后顺序进行有监督学习，后一次有监督学习在前一次有监督学习得到的改进深度学习模型上进行，从而使用时间越靠后越接近将来使用的时间，即越是后输入的数据越接近将来的使用时间，而越

后输入的数据是过去最近的时间产生的数据，从而使得最终进化后的深度学习模型更能适应将来使用的需要，准确率更高。

在一个技术方案中，上述深度学习模型进化装置包括清零模块（图 2-8 中未说明），用于删除所有数据组，将进化后的深度学习模型作为新的已训练的深度学习模型，并返回数据获取模块 1 重新调取已训练的深度学习模型，将输入数据输入已训练的深度学习模型进行深度学习，得到输出数据并记录使用时间及获取对应输入数据的真实结果数据。如此，可进行清零，避免存储的数据组对下一轮进化产生影响。

为更好地说明上述深度学习模型进化方法和装置的处理过程，以下以一个具体技术方案进行举例说明。

预设时间段为从最早的使用时间到现在为止；预设阈值为 0；组数阈值为 3。

比较进化知识库中每组数据组中输出数据与对应的真实结果数据的差异值，统计从最早的使用时间到现在为止差异值大于 0 的组数，如果组数大于 3，则利用进化知识库中从最早的使用时间到现在为止差异值大于 0 的数据组中输入数据和对应的真实结果数据对当前已训练的深度学习模型进行训练。

输入数据人脸图像记为 IM，输出数据记为 OD，对应的真实结果数据记为 TD（为 1 时代表是张三人脸，为 0 时代表不是张三人脸），输出数据与对应的真实结果数据的差异值记为 DD，从最早的使用时间到现在为止差异值大于 0 的组数记为 GS。

进化知识库中有如下数据组：

1 月 1 日，IM＝A，OD＝1，TD＝1，DD＝0，GS＝0。

2 月 1 日，IM＝B，OD＝1，TD＝1，DD＝0，GS＝0。

3 月 1 日，IM＝C，OD＝0，TD＝0，DD＝0，GS＝0。

4 月 1 日，IM＝D，OD＝0，TD＝0，DD＝0，GS＝0。

5 月 1 日，IM＝E，OD＝1，TD＝0，DD＞0，GS＝1。

6 月 1 日，IM＝F，OD＝0，TD＝0，DD＝0，GS＝1。

7 月 1 日，IM＝G，OD＝1，TD＝1，DD＝0，GS＝1。

8 月 1 日，IM＝H，OD＝0，TD＝1，DD＞0，GS＝2。

9 月 1 日，IM＝I，OD＝1，TD＝1，DD＝0，GS＝2。

10 月 1 日，IM＝J，OD＝0，TD＝1，DD＞0，GS＝3。

11 月 1 日，IM＝K，OD＝1，TD＝0，DD＞0，GS＝4。

此时从最早的使用时间到现在为止差异值大于 0 的组数 4 大于 3，则利用进化知识库中从最早的使用时间到现在为止差异值大于 0 的数据组中输入数据和对应的真实结果数据对当前已训练的深度学习模型进行训练。

初始的深度学习模型记为 M0。先进行无监督学习，将输入数据人脸图像（记为 IM）输入当前已训练的深度学习模型（记为 PM），对当前已训练的深度学习模型进行无监督学习，得到训练深度学习模型（记为 CM）：

5 月 1 日，IM＝E。PM＝M0，CM＝M0501。

8 月 1 日，IM＝H。PM＝M0501，CM＝M0801。

10 月 1 日，IM＝J。PM＝M0801，CM＝M1001。

11 月 1 日，IM＝K。PM＝M1001，CM＝M1101。

然后进行有监督学习，将输入数据人脸图像（记为 IM）、对应的真实结果数据（记为TD）作为初步进化的深度学习模型的输入和期望输出，对初步进化的深度学习模型（记为PN）进行有监督学习，得到改进深度学习模型（记为 CN）：

5 月 1 日，IM＝E，TD＝0。PN＝M1001，CN＝N0501。

8 月 1 日，IM＝H，TD＝1。PN＝N0501，CN＝N0801。

10 月 1 日，IM＝J，TD＝1。PN＝N0801，CN＝N1001。

11 月 1 日，IM＝K，TD＝0。PN＝N1001，CN＝N1101。



第3章

适应型深度学习模型

深度学习模型不是万能的，在有些场景或应用中会出现局限性，因此需要提高现有深度学习模型的适应性。本章给出了两种适应型深度学习模型：期望与反期望深度学习模型、残缺数据深度学习模型。

3.1 期望与反期望深度学习模型

深度学习模型通过输入数据和输出数据对深度学习模型进行训练。如果有的输入数据在不同场景下能产生属性相反的输出数据（这是有可能的，假如还有影响输出结果的其他条件数据没有包括在输入数据中，那么就有可能因为其他条件数据的变化，而产生不同的输出数据），此时就会在训练过程中引起深度学习模型结构的动荡，不利于深度学习模型的训练，也不利于深度学习模型的使用，因为无法得知输出数据的可信度是多少[11,12]。

例如，深度学习模型判断一个人是不是男的，输入不同人的人脸特征，同样脸部特征的人可能是男的，也可能是女的，此时就会在训练过程中引起深度学习模型结构的动荡。

本技术研究并给出了一种新的期望与反期望深度学习方法和神经网络系统，通过输出数据的期望标签和反期望标签构造两个深度学习模型，分别对与期望标签对应的深度学习模型和与反期望标签对应的深度学习模型进行训练，避免了深度学习模型当存在属性相反的输出数据时结构动荡的问题，提高了深度学习模型训练的可靠性。

3.1.1 期望与反期望深度学习方法

如图 3-1 所示,本技术方案提供一种期望与反期望深度学习方法,可包括以下步骤:

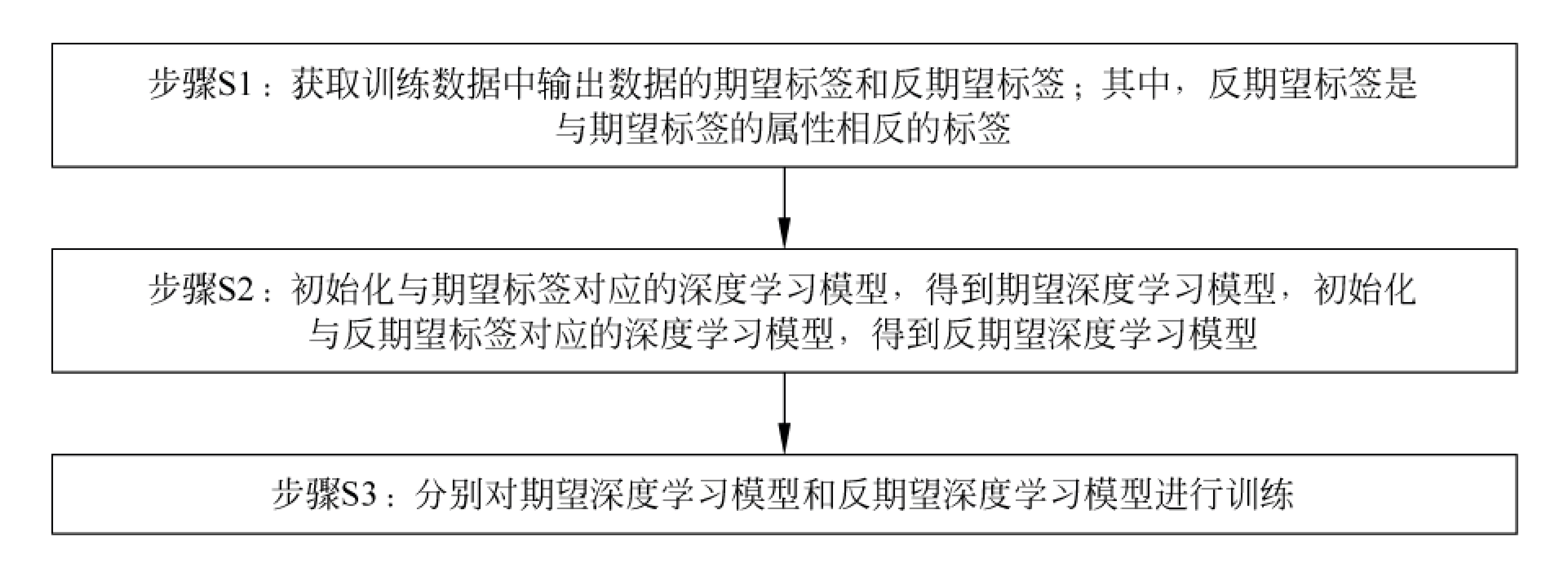


图 3-1 期望与反期望深度学习方法流程

(1) 步骤 S1,获取训练数据中输出数据的期望标签和反期望标签;其中,反期望标签是与期望标签的属性相反的标签。

例如,获取输出数据的期望标签“男人”,反期望标签“女人”。

(2) 步骤 S2,初始化与期望标签对应的深度学习模型,得到期望深度学习模型;初始化与反期望标签对应的深度学习模型,得到反期望深度学习模型。

其中,期望标签对应的深度学习模型的输入格式初始化为训练数据中的输入数据格式;反期望标签对应的深度学习模型的输入格式也初始化为训练数据中的输入数据格式。期望深度学习模型的输出格式初始化为 0~1 之间的数,输出为 1 时,表示输出为期望标签;输出为 0 时,表示输出不是期望标签;输出为 0~1 时,表示输出有可能是期望标签。反期望深度学习模型的输出格式初始化为 0~1 的数,输出为 1 时,表示输出为反期望标签;输出为 0 时,表示输出不是反期望标签;输出为 0~1 时,表示输出有可能是反期望标签。同时,获取已有同类深度学习模型的配置信息(预设配置信息包括:预设层数、每层预设节点数、各网络连接的预设权值)作为期望标签对应的深度学习模型的配置信息对期望标签对应的深度学习模型进行配置,同时也将该配置信息作为反期望标签对应的深度学习模型的配置信息对反期望标签对应的深度学习模型进行配置。

例如,初始化与期望标签“男人”和反期望标签“女人”对应的两个深度学习模型,称为期望标签“男人”深度学习模型、反期望标签“女人”深度学习模型。

(3) 步骤 S3,分别对期望深度学习模型和反期望深度学习模型进行训练。

在一个技术方案中,本步骤可以采用以下方式实现:

① 步骤 S3-1,通过训练数据中的输入数据分别对期望深度学习模型和反期望深度学习模型进行无监督训练。

例如,通过训练数据中的输入数据人脸图像分别对期望标签“男人”深度学习模型、反期望标签“女人”深度学习模型进行无监督训练。对期望深度学习模型、反期望深度学习模型进行无监督训练,可采用同样的输入数据集合。

② 步骤S3-2，从训练数据中获取与期望标签一致的输出标签对应的输入数据，将该输入数据作为输入，将“1”作为预期输出，对期望深度学习模型进行有监督训练，并从训练数据中获取与期望标签及反期望标签都不一致的输出标签对应的输入数据，将该输入数据作为输入，将“0”作为预期输出，对期望深度学习模型进行有监督训练。

可以将训练数据中与期望标签一致的输出标签对应的输入数据作为第一输入数据；从第一输入数据中筛选出第一预设比例(假设为$P\%$，P为0～100的实数)的第一输入数据，得到对应的每一第二输入数据(即将第一输入数据中$P\%$的数据保留，其余数据清空，得到对应的每一第二输入数据)；将每一第二输入数据作为输入，将第一预设比例作为预期输出，对期望深度学习模型进行有监督训练。

之所以不进行从训练数据中获取与反期望标签一致的输出标签对应的输入数据，将该输入数据作为输入，将“0”作为预期输出，对期望深度学习模型进行有监督训练，是因为与反期望标签一致的输出标签所对应的输入数据有可能与期望标签一致的输出标签所对应的输入数据类似(例如，有的男人和女人的脸部特征很相像)，从而导致类似的输入数据在同一个深度学习模型中产生不同的输出标签，进而会对期望深度学习模型的训练效果产生负面影响。

例如，从训练数据中获取与期望标签“男人”一致的输出标签(预期输出为“1”)及其对应输入数据人脸图像对期望标签“男人”深度学习模型进行有监督训练；从训练数据中获取与期望标签“男人”及反期望标签“女人”都不一致的输出标签(预期输出为“0”)及其对应输入数据(例如动物头像)对期望标签“男人”深度学习模型进行有监督训练；从训练数据中获取与期望标签“男人”一致的输出标签(预期输出为“1”)及其对应输入数据人脸图像，将输入数据人脸图像中60%的部分保留(可随机选择，也可以均匀选择，或按照一定预设方式选择需要保留的部分)，其余40%的部分从图像中清除(可随机选择，也可以均匀选择，或按照一定预设方式选择需要保留的部分)，然后将处理后的输入数据人脸图像作为新输入数据人脸图像，将输出数据“1”变为新输出数据“60%”，对期望标签“男人”深度学习模型进行有监督训练。

③ 步骤S3-3，从训练数据中获取与反期望标签一致的输出标签对应的输入数据，将该输入数据作为输入，将“1”作为预期输出，对反期望深度学习模型进行有监督训练，并从训练数据中获取与期望标签及反期望标签都不一致的输出标签对应的输入数据，将该输入数据作为输入，将“0”作为预期输出，对反期望深度学习模型进行有监督训练。

可以将训练数据中与反期望标签一致的输出标签对应的输入数据作为第三输入数据；从每一第三输入数据中筛选出第二预设比例(假设为$P\%$，P为0～100的实数)的数据，得到对应的每一第四输入数据(即将每一第三输入数据中$P\%$的数据保留，其余数据清空，得到对应的每一第四输入数据)；将每一第四输入数据作为输入，将对应的第二预设比例作为预期输出，对反期望深度学习模型进行有监督训练。

例如，从训练数据中获取与反期望标签“女人”一致的输出标签(预期输出为“1”)及其对应输入数据人脸图像对反期望标签“女人”深度学习模型进行有监督训练；从训练数据中获取与期望标签“男人”及反期望标签“女人”都不一致的输出标签(预期输出为“0”)及其对应输入数据(例如动物头像)对期望标签“女人”深度学习模型进行有监督训练。

从训练数据中获取与反期望标签“女人”一致的输出标签(预期输出为“1”)及其对应输

入数据人脸图像，将输入数据人脸图像中60%的部分保留(可随机选择，也可以均匀选择，或按照一定预设方式选择需要保留的部分)，其余40%的部分从图像中清除(可随机选择，也可以均匀选择，或按照一定预设方式选择需要保留的部分)，然后将处理后的输入数据人脸图像作为新输入数据人脸图像，将输出数据“1”变为新输出数据“60%”，对反期望标签“女人”深度学习模型进行有监督训练。

在分别对期望深度学习模型和反期望深度学习模型进行训练之后，还可以计算输入数据对应的输出属于期望标签及反期望标签的可信度。可以将输入数据输入期望深度学习模型，得到期望深度学习模型的输出数据；将输入数据输入反期望深度学习模型，得到反期望深度学习模型的输出数据；根据期望深度学习模型与反期望深度学习模型的输出数据得到输入数据对应的输出属于期望标签及反期望标签的可信度。

例如，获取输入数据人脸图像，将输入数据人脸图像输入期望深度学习模型，得到期望深度学习模型的输出数据，该输出数据为0～1的数，越接近1，则表明输出为期望标签的概率越大；将输入数据输入反期望深度学习模型，得到反期望深度学习模型的输出数据，该输出数据为0～1的数，越接近1，则表明输出为反期望标签的概率越大。

在计算可信度时，可以将期望深度学习模型的输出数据作为输出属于期望标签的概率记为 A，将反期望深度学习模型的输出数据作为输出属于反期望标签的概率记为 B；将输入数据对应的输出属于期望标签或反期望标签的可信度记为 F，则 $F=(A+B)/2$；输入数据对应的输出属于期望标签的可信度为 $F\times A$，属于相反标签的可信度为 $F\times B$，属于其他标签的可信度为 $1-F$。

例如，如果这个人是男人，那么这个人是男人的概率是 $F\times A$，作为这个人是男人的可信度；是女人的概率是 $F\times B$，作为这个人是女人的可信度；属于其他标签的概率为 $1-F$，作为这个人既不是男人又不是女人的可信度。

深度学习模型判断一个人是不是男人，训练数据中输入头像图像，如果训练数据中输出数据是男人，则用“男人”标签对应的深度学习模型进行训练，例如用输出“1”代表是男人，用“0”代表不是男人(例如是动物)；如果训练数据中输出数据是女人，则用“女人”标签对应的深度学习模型进行训练，例如用输出用0～1的数来代表是女人的可能性大小，用1代表是女人，用0代表不是女人(例如是动物)；如果训练数据中输出数据是跟男人、女人无关的，例如动物，则“男人”标签对应的深度学习模型和“女人”标签对应的深度学习模型同时进行训练。

在使用时，输入一个数据到“男人”和“女人”标签对应的深度学习模型同时计算，如果“男人”标签对应的深度学习模型输出0.8；如果“女人”标签对应的深度学习模型输出0.5，则输出是“男人”或“女人”的可信度是(0.8+0.5)/2=0.65=65%，输出不是“男人”或“女人”的可信度是1−65%=35%，输出是“男人”的可信度是65%×0.8=52%，输出是“女人”的可信度是65%×0.5=32.5%，相对而言是“男人”的可信度更高，所以判断为“男人”。

3.1.2 期望与反期望深度学习系统

如图3-2所示，本技术方案还提供一种神经网络系统，可包括：

(1) 标签获取模块1，用于获取训练数据中输出数据的期望标签和反期望标签；其中，

反期望标签是与期望标签的属性相反的标签。

例如，获取输出数据的期望标签“男人”和反期望标签“女人”。

（2）初始化模块 2，用于初始化与期望标签对应的深度学习模型，得到期望深度学习模型，初始化与反期望标签对应的深度学习模型，得到反期望深度学习模型。

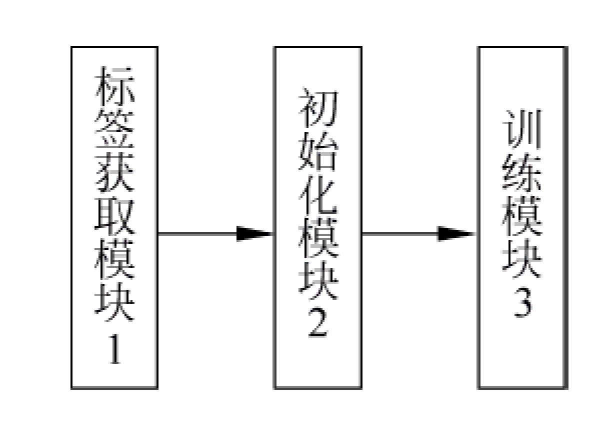


图 3-2　深度学习系统的结构示意图

其中，期望标签对应的深度学习模型的输入格式初始化为训练数据中的输入数据格式；反期望标签对应的深度学习模型的输入格式也初始化为训练数据中的输入数据格式。期望深度学习模型的输出格式初始化为 0～1 的数，输出为“1”时，表示输出为期望标签；输出为“0”时，表示输出不是期望标签；输出为 0～1 的数时，表示输出有可能是期望标签。反期望深度学习模型的输出格式初始化为 0～1 的数，输出为“1”时，表示输出为反期望标签，输出为“0”时，表示输出不是反期望标签；输出为 0～1 的数时，表示输出有可能是反期望标签。同时，获取已有同类深度学习模型的配置信息（预设配置信息包括：预设层数、每层预设节点数、各网络连接的预设权值）作为期望标签对应的深度学习模型的配置信息对期望标签对应的深度学习模型进行配置，同时也将该配置信息作为反期望标签对应的深度学习模型的配置信息对反期望标签对应的深度学习模型进行配置。

例如，初始化与期望标签“男人”和反期望标签“女人”对应的两个深度学习模型，称为期望标签“男人”深度学习模型、反期望标签“女人”深度学习模型。

（3）训练模块 3，用于分别对期望深度学习模型和反期望深度学习模型进行训练。

本技术方案通过输出数据的期望标签和反期望标签构造两个深度学习模型，如果训练数据中输出数据与期望标签一致，则对与期望标签对应的深度学习模型进行训练；如果训练数据中输出数据与反期望标签一致，则对与反期望标签对应的深度学习模型进行训练；如果训练数据中输出数据与属性无关（当然也与反期望标签无关），则同时对期望标签及反期望标签对应的两个深度学习模型进行训练。训练得到两个深度学习模型，在应用中将输入数据同时输入这两个深度学习模型，得到两个输出数据，然后综合这两个输出数据得到输出数据属于期望标签及反期望标签的可信度。

本技术方案的神经网络系统与期望和反期望深度学习方法一一对应，上述期望与反期望深度学习方法的技术方案阐述的技术特征及其有益效果均适用于神经网络系统的技术方案中。

3.2　残缺深度学习模型

人工神经网络（Artificial Neural Network，ANN）也简称为神经网络或连接模型，它是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间的相互连接关系，从而达到处理信息的目的。

传统技术中，神经网络包括用于深度学习的神经网络都是通过输入数据和输出数据对神经网络进行训练，然后在使用阶段将输入数据输入已经训练好的神经网络，从而得到输出

数据。但是,在实际使用过程中,输入数据往往都不完整,因此基于传统技术的神经网络无法根据不完整的输入数据得到输出数据[13,14]。

本技术方案研究并给出了一种新型的残缺数据深度学习模型方法、系统,方法包括:获取预先训练好的深度学习模型;从待预测数据中获取输入数据;根据输入数据和预先训练好的深度学习模型的输入变量确定至少一个缺少数据的输入变量;将对缺少数据的输入变量进行赋值获得的数据,与从待预测数据中获取的输入数据进行组合,生成多组完整输入数据;将多组完整输入数据输入到预先训练好的深度学习模型中,计算得到多组输出数据。通过这种方式,在输入数据缺少数据的情况下,仍然能够通过神经网络输出预测的输出变量,为研究工作提供数据支持,减少了输入数据检测和采集的成本。

3.2.1　残缺深度学习模型方法

1. 残缺深度学习模型的基本方法

如图3-3所示,本技术方案提供了一种深度学习模型方法,该方法以应用于服务器进行举例说明。包括:

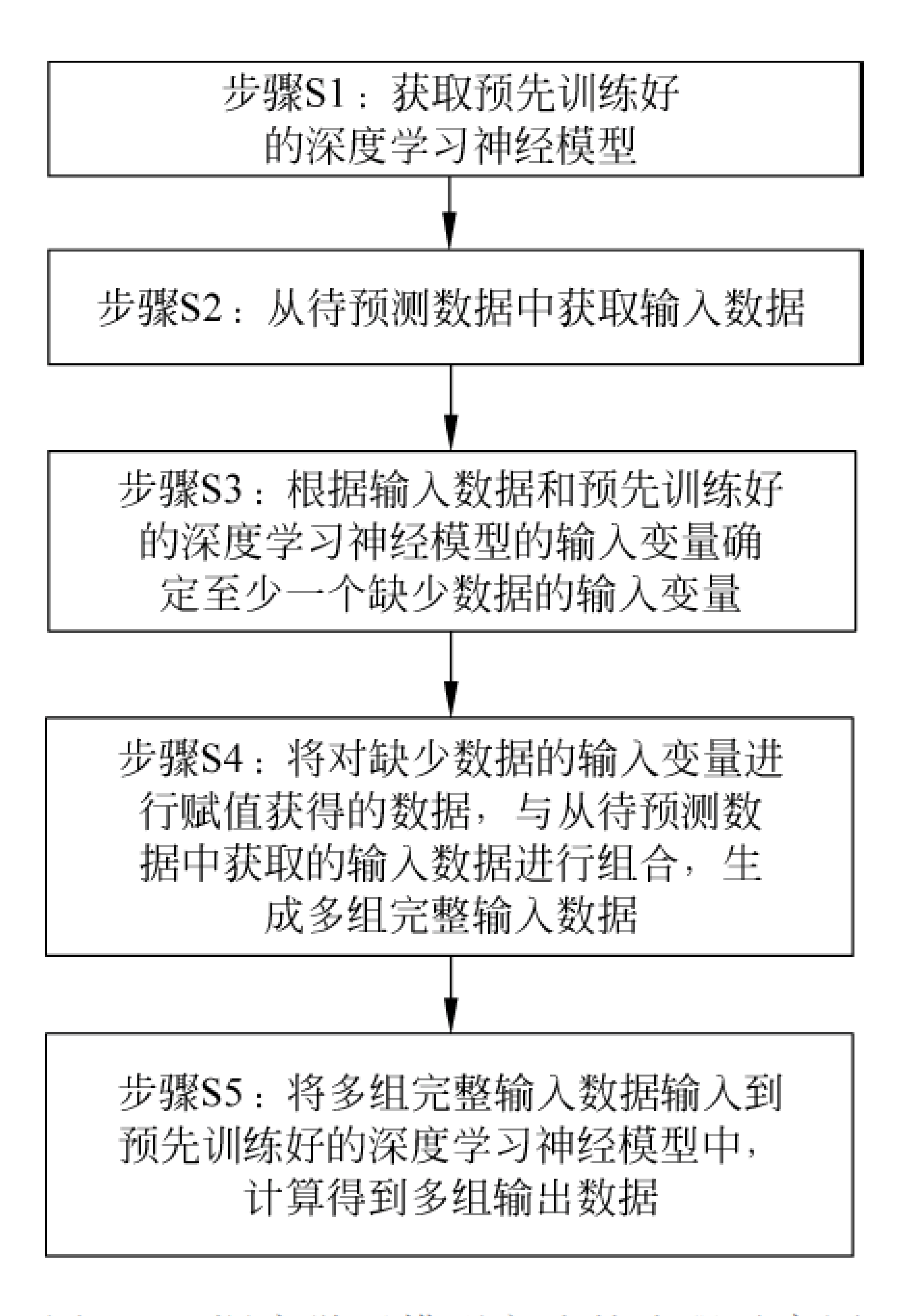


图3-3　深度学习模型方法的流程示意图

(1) 步骤S1,获取预先训练好的深度学习模型。

(2) 步骤S2,从待预测数据中获取输入数据。

(3) 步骤S3,根据输入数据和预先训练好的深度学习模型的输入变量确定至少一个缺少数据的输入变量。

预先训练好的深度学习模型有 N 个输入变量。待预测数据中有其中 M 个输入变量对应的数据,其余 $N-M$ 个输入变量没有对应的数据,则这 $N-M$ 个输入变量是缺少数据的

输入变量；其中，M、N 为自然数。

例如，深度学习模型输入变量为：身高、体重、血压、血液检查结果、尿液检查结果；深度学习模型输出是糖尿病类型(例如，用 0、1、2、3 来表示不同类型的糖尿病)。

待测试的不完整输入数据包括血压数据、血液检查结果数据、尿液检查结果数据，缺少数据输入变量为身高、体重。

(4) 步骤 S4，将对缺少数据的输入变量进行赋值获得的数据，与从待预测数据中获取的输入数据进行组合，生成多组完整输入数据。

通过知识库获取缺少数据的输入变量中每一缺少数据的输入变量的取值范围。具体包括 3 种方式：

第 1 种方式：从输入变量范围知识库中检索出缺少数据输入变量对应的预设取值范围；其中，输入变量范围知识库预先存有所有输入变量对应的预设取值范围。

第 2 种方式：提示并接受用户输入缺少数据输入变量的预设取值范围。

第 3 种方式：从每一训练输入数据中获取缺少数据输入变量的值，统计获取到的缺少数据输入变量的最大值和最小值，将最小值到最大值的范围作为缺少数据输入变量的预设取值范围。

例如，身高的取值范围是 0.1～2m，体重的取值范围是 2～100kg。

知识库中存储有多组数据，其中包括多组预先训练好的深度学习模型输入数据。可通过知识库中存储的数据获取缺少数据输入数据的取值范围。

对缺少数据输入变量在对应的预设取值范围内进行 N 次赋值，具体包括两种实现方法：

第 1 种实现方法：对缺少数据输入变量在对应的预设取值范围内随机地取 N 个值对缺少数据输入变量进行 N 次赋值。

第 2 种实现方法：从所有训练输入数据中随机抽取 N 个训练输入数据，然后提取其中该输入变量对应的数据对缺少数据输入变量进行 N 次随机赋值。

其中，预设次数 N 的确定方法包括：

接受用户输入预设次数 N 的值，如果用户没有输入，则将训练输入数据的数量作为 N 的值。

例如，对缺少数据输入变量身高、体重在对应的预设取值范围 0.1～2m、2～100kg 内复制其他样本中该输入变量数据进行 100 次随机赋值，因为在很多样本中是具备身高、体重数据的，所以通过获取已有其他样本中这些输入变量的数据是可行的：

身高 0.5m，体重 10kg；

身高 1.5m，体重 50kg；

……

身高 1.7m，体重 55kg；

将该缺少数据输入变量身高、体重对应的 100 份输入数据与待测试的不完整输入数据"血压数据 C、血液检查结果数据 D、尿液检查结果数据 E"组合得到 100 份完整的输入数据：

身高 0.5m，体重 10kg，血压数据 C，血液检查结果数据 D，尿液检查结果数据 E；

身高 1.5m，体重 50kg，血压数据 C，血液检查结果数据 D，尿液检查结果数据 E；

……

身高 1.7m，体重 55kg，血压数据 C，血液检查结果数据 D，尿液检查结果数据 E。

通过知识库中存储的数据获取缺少输入数据的取值范围后，在此取值范围内取预定数量的数据，将选取的数据赋值给缺少数据的输入变量。预定数量是预先设置的数量值，可根据不同的输入变量设定不同的预定数量。选择 M 个数量的数据对缺少数据的输入变量进行赋值，就能得到 M 个缺少数据的输入变量的输入数据，再将填入了数据的缺少数据的输入变量与从待预测数据中获取的输入数据进行组合，生成多组完整的输入数据。

(5) 步骤 S5，将多组完整输入数据输入到预先训练好的深度学习模型中，计算得到多组输出数据。

将缺少数据的输入变量缺少数据的输入数据补充完整后，与从待预测数据中获取的输入数据进行组合，生成多组完整的输入数据。再将多组完整的输入数据输入到预先训练好的深度学习模型中，则能够获得多个输出数据。

例如，将 100 份完整的输入数据输入已训练的深度学习模型，得到多份输出：

身高 0.5m，体重 10kg，血压 C，血液检查结果 D，尿液检查结果 E；该份完整的输入数据输入已训练的深度学习模型得到输出 Y_1；

身高 1.5m，体重 50kg，血压 C，血液检查结果 D，尿液检查结果 E；该份完整的输入数据输入已训练的深度学习模型得到输出 Y_2；

……

身高 1.7m，体重 55kg，血压 C，血液检查结果 D，尿液检查结果 E；该份完整的输入数据输入已训练的深度学习模型得到输出 Y_{100}。

通过这种方式，在输入数据缺少数据的情况下，仍然能够通过神经网络输出预测的输出变量，为研究工作提供数据支持。

2. 残缺深度学习模型的详细方法

在一个技术方案中，将对缺少数据的输入变量进行赋值获得的数据，与从待预测数据中获取的输入数据进行组合，生成多组完整输入数据，包括：从输入变量范围知识库中检索出缺少数据输入变量对应的预设取值范围，其中，输入变量范围知识库预先存有所有输入变量对应的预设取值范围；或从每一训练输入数据中获取缺少数据输入变量的值，统计获取缺少数据输入变量的最大值和最小值，将从最小值到最大值之间值的范围作为缺少数据输入变量的预设取值范围；或提示并接受用户输入缺少数据输入变量的预设取值范围，对缺少数据输入变量在对应的预设取值范围内随机地生成预定数量个值对缺少数据的输入变量进行预定数量次赋值，其中，预定数量由用户输入或预先设置。

当待预测数据中缺少一个或多个输入变量的数据时，则需要从知识库存储的对应数据中选取预定数量的数据对缺少数据的输入变量进行赋值。预定数量可以是用户通过输入设备输入的，当用户未输入时，预定数量可以根据预设比例计算。具体而言，先根据缺少数据的输入变量从知识库中查找缺少数据输入数据数量，例如缺少数据的输入变量为体重，那么就从知识库中查找体重这一项的数据，当体重这一项的数据有 120 项时，那么知识库中缺少数据的输入变量为体重的对应数据数量则为 120。再获取到预设比例为 2/3，那么预定数量则等于 120 乘以 2/3，即预定数量为 80。当计算出来的预定数量不是整数时，则可以通过四舍五入或取整的方式取整数值。预设比例的设定可由研究人员根据研究需求而定。预设比例的设定较灵活，可根据实际项目需求进行调整，根据预设比例与总体数据数量进行预设数

量的确定，也能保证选取数据的数量，提高输出数据的总体准确性。

在一个技术方案中，在将多组完整输入数据输入到预先训练好的深度学习模型中，计算得到多组输出数据之后，统计多组输出数据中各类输出数据所占的比例；将比例最高的一类输出数据作为预测结果；将预测结果及各类输出数据及其相应比例输出给用户。

在得到多个输出数据之后，需要分别统计各类输出数据所占的比例，再将比例最高的输出数据作为预测结果。

例如，输入变量为体重、身高、血压、血液检查结果，输出变量为糖尿病严重程度等，当输入多组体重、身高、血压、血液检查结果的数据后，能获得多组输出的关于糖尿病严重程度的数据。当输出的糖尿病严重程度数据为"0"的有 10 个，为"1"的有 30 个，为"2"的有 15 个时，则将占比例最高的为"1"的数据作为预测结果。其中，数据为"0"代表没有糖尿病，数据为"1"代表有轻微糖尿病，数据为"2"代表有严重糖尿病。具体的数据输出可由研究人员进行自定义设置。

例如，根据 Y_1、Y_2、…、Y_{100} 的值进行统计，其中输出为 0 的有 10 个，为"1"的有 30 个，为"2"的有 60 个。其中"0"代表没有糖尿病，"1"代表有轻微糖尿病，"2"代表有严重糖尿病，则输出为"0"(代表没有糖尿病)的比例是 10%，输出为"1"(代表轻微糖尿病)的比例是 30%，输出为"2"(代表严重糖尿病)的比例是 60%。

将多个输出数据及其对应比例输出给用户，作为不完整输入数据的输出数据；将最高比例对应的输出数据输出给用户，作为不完整输入数据的最大概率输出数据。

例如，将输出为"0"(代表没有糖尿病)的比例是 10%，输出为"1"(代表轻微糖尿病)的比例是 30%，输出为"2"(代表严重糖尿病)的比例是 60%输出给用户，作为不完整输入数据血压 C、血液检查结果 D、尿液检查结果 E 的输出。

在医疗行业，虽然有诊疗神经网络模型，但输入变量很多，例如糖尿病神经网络模型，输出是无糖尿病、X 类糖尿病、Y 类糖尿病等(实际中 X、Y 有具体名称)，输入为血压、身高、体重、血液检查结果(实际中包括多项具体指标)、尿液检查结果(实际中包括多项具体指标)。但患者刚开始来的时候，可能其中血液没有检查，那么在现有技术中就无法使用该模型了，而采用本技术方案可以通过已有的患者输入数据，大概地判别各种疾病的可能性，进行预诊和分诊，假如通过现有残缺的数据就足以判断出想要的结果，例如，现有数据输入后，发现输出的结果只有无糖尿病这一类，那么就无须再多做检查，反而节省了输入数据检测和采集的成本；或者看到输出为"2"(代表严重糖尿病)的比例是 60%，说明患严重糖尿病的可能性最大，则需要进一步做更深入的检查。

当输出数据有多个时，则需要对输出结果进行筛选，将比例较高的数据作为预测输出数据，更能确保预测数据的准确性。

在一个技术方案中，将对缺少数据的输入变量进行赋值获得的数据，与从待预测数据中获取的输入数据进行组合，生成多组完整输入数据，包括：从样本数据库中随机选取预定数量的训练输入数据，然后提取这些训练输入数据中该输入变量对应的数据对缺少数据输入变量进行预定数量次赋值；其中，预定数量由用户输入或将小于等于样本数据库中输入数据的数量的 P% 作为预定数量，P 小于等于 100，样本数据库中的输入数据都是完整的输入数据。

在一个技术方案中，在从待预测数据中获取输入变量之前，对深度学习模型进行初始

化；获取训练数据；通过训练数据对初始化后的深度学习模型进行训练得到预先训练好的深度学习模型。

深度学习模型在根据输入数据预测输出数据之前，需要先获取深度学习模型并对其进行初始化，再获取训练数据，通过训练数据对初始化后的深度学习模型进行训练，训练后则能获得训练好的深度学习模型，训练好的深度学习模型才能用于对输出数据的预测操作。深度学习模型可以自行创建，也可以为了节约创建时间，直接从深度学习模型库中获取匹配度较高的深度学习模型，再对获取到的深度学习模型进行初始化，初始化后的深度学习模型才能与当前研究项目需要的输入变量与输出变量更契合，而对深度学习模型训练后才能使得深度学习模型更准确地预测输出数据。

在一个技术方案中，对深度学习模型进行初始化，包括：根据样本数据库中训练数据的字段名确定样本输入变量与样本输出变量；从深度学习模型库中选取输入变量和输出变量与样本输入变量和样本输出变量匹配的深度学习模型；对选取的深度学习模型进行初始化。

在知识库中存储有多种类型的数据，每个类型的数据有对应的字段。例如，存储的数据中包括有体重、身高、血压这三种类型的数据，那么对应的字段可设为体重、身高、血压。再根据研究项目需求，从知识库中选择需要用到的字段，将选取的字段确认为样本输入变量与样本输出变量。例如，当项目需要研究根据身高、体重、血压、血液检测结果等数据预测该用户是否患有糖尿病和患病严重程度时，可从知识库中选择体重、血压、血液检测结果的字段作为样本输入变量，并选择糖尿病类型作为样本输出变量，则项目研究需要的深度学习模型的输入变量为体重、血压、血液检测结果，而输出变量则为糖尿病类型。

当确定了样本输入变量与样本输出变量后，可根据样本输入变量与样本输出变量从深度学习模型库中选取相匹配的深度学习模型。由于深度学习模型库中的神经网络模型数量较多，在选取与实际需求对应的神经网络模型时，应该先根据研究需求从知识库中选择对应的字段将项目需要的神经网络模型的输入变量与输出变量设定，然后再根据设定的输入变量与输出变量从深度学习模型库中查找与实际项目需求相匹配的神经网络模型。这样能减少选择神经网络模型的操作步骤，节约时间成本。

在一个技术方案中，从深度学习模型库中选取输入变量和输出变量与样本输入变量和样本输出变量匹配的深度学习模型，采用如下公式计算深度学习模型库中输入变量和输出变量分别与样本输入变量和样本输出变量之间的匹配度：

$$\text{匹配度}=\text{深度学习模型库中输入变量与样本输入变量的匹配度}\times u\%+\text{深度学习模型库中输出变量与样本输出变量的匹配度}\times(1-u\%)$$

其中，u 为预设权重。从匹配度中选取匹配度最大的深度学习模型返回给用户；将用户确认或修改后的深度学习模型作为预置深度学习模型。

当根据样本输入变量和样本输出变量从深度学习模型库中选择匹配的深度学习模型时，可将深度学习模型库中的多个深度学习模型的输入变量和输出变量分别与样本输入变量和样本输出变量进行对比匹配，再将两者的匹配度综合，选择综合匹配度最高的。具体而言，可以设置一个预定权重，例如设定权重为 50 时，则说明研究人员认为输入变量与输出变量的重要性是一致的。将深度学习模型库中的多个深度学习模型的输入变量与样本输入变量一一对比，获取输入变量与样本输入变量的匹配度。再将深度学习模型库中的多个深度

学习模型的输出变量与样本输出变量一一对比,获取输出变量与样本输出变量的匹配度。

例如,神经网络 A 的输入变量与样本输入变量的匹配度为 4/5,神经网络 A 的输出变量与样本输出变量的匹配度为 3/4,那么神经网络 A 的综合匹配度为 4/5×50%+3/4×(1−50%)=0.775。当将深度学习模型库中的多个深度学习模型都一一匹配完毕后,统计每个深度学习模型与样本输入变量和样本输出变量的综合匹配度,并将匹配度最高的神经网络作为预置深度学习模型。因此,若是其他神经网络的综合匹配度低于 0.775 时,则选择匹配度最高的神经网络 A 作为预置深度学习模型。

将样本输入变量与样本输出变量均作为挑选标准,能在更大程度上保证“挑选”出的深度学习模型与实际项目需求的契合度,更能节约研究的时间成本。

在一个技术方案中,对初始化后的深度学习模型进行训练得到预先训练好的深度学习模型,包括:获取初始化后的深度学习模型中的输入变量与输出变量;从知识库中获取输入数据,将其输入到深度学习模型中进行无监督训练;从知识库中获取输入数据及输出数据,将输入数据作为初始化后的深度学习模型的输入,将输出数据作为初始化后的深度学习模型的预期输出,对初始化后的深度学习模型进行有监督训练。将每个训练输入数据作为深度学习模型的输入,对深度学习模型进行无监督训练;将每个带标签的训练输入数据及其对应的预期标签作为深度学习模型的输入和预期输出,对深度学习模型进行有监督训练。

当从深度学习模型库中选择了某一个深度学习模型后,会先对该深度学习模型进行初始化,得到初始化后的深度学习模型,初始化后的深度学习模型对应有输入变量与输出变量。然后对初始化后的深度学习模型进行训练,训练分为无监督训练和有监督训练。一般情况下,先进行无监督训练,再进行有监督训练。无监督训练是指从知识库中获取输入数据,将输入数据输入到初始化后的深度学习模型中进行无监督训练。无监督训练不需要预期输出,而有监督训练则需要预期输出。

例如,当输入变量为体重、身高、血压、血液检测结果、尿液检测结果时,从知识库中获取多组体重、身高、血压、血液检测结果、尿液检测结果字段对应的数据作为输入变量对应的数据,可以对初始化后的深度学习模型进行无监督训练。

在无监督训练后,再对深度学习模型进行有监督训练。进行有监督训练时,会将深度学习模型的输入数据与预期输出均提供。例如,当体重、身高、血压、血液检测结果、尿液检测结果为输入变量,糖尿病类型为输出变量时,从知识库中获取体重、身高、血压、血液检测结果、尿液检测结果字段对应的多组数据作为输入变量对应的数据,将糖尿病类型字段对应的数据作为输出变量对应的数据,对始化后的深度学习模型进行有监督训练。

用户样本数据中存储有多组用户数据。例如,每组用户数据中的数据类型包括有每个用户的体重、身高、血压、血液检测结果、尿液检测结果以及糖尿病类型数据。但是,并不是每组用户数据都是完整的,可能存在有的用户数据中缺少糖尿病类型数据,这部分缺少糖尿病类型数据的用户数据可以用于对深度学习模型进行无监督训练,避免了数据的浪费。由于无监督训练先训练了深度学习模型的特征提取能力,在进行了无监督训练后再进行有监督训练也能提升深度学习模型的训练效果,提升训练后的深度学习模型预测输出数据的准确度。

3.2.2　残缺深度学习模型系统

图 3-4 提供了一种深度学习模型装置，包括：

(1) 数据获取模块 1，用于获取预先训练好的深度学习模型，从待预测数据中获取输入数据。

(2) 缺少数据的输入变量确认模块 2，用于根据输入数据和预先训练好的深度学习模型的输入变量确定至少一个缺少数据的输入变量。

(3) 取值范围获取模块 3，用于将对缺少数据的输入变量进行赋值获得的数据，与从待预测数据中获取的输入数据进行组合，生成多组完整输入数据。

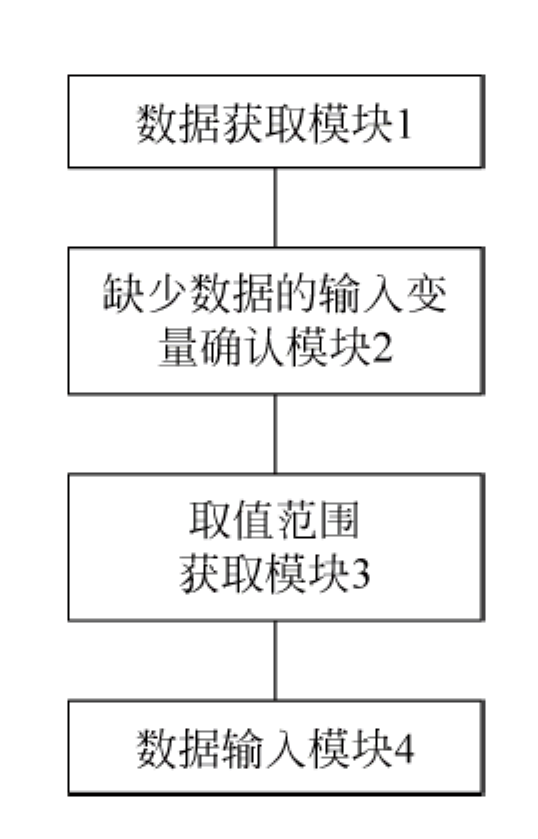


图 3-4　深度学习模型方法的装置结构框图

在一个技术方案中，取值范围获取模块 3 还用于从输入变量范围知识库中检索出缺少数据输入变量对应的预设取值范围，其中，输入变量范围知识库预先存有所有输入变量对应的预设取值范围；或从所有训练输入数据中获取缺少数据输入变量的值，统计获取到的缺少数据输入变量的最大值和最小值，将从最小值到最大值之间值的范围作为缺少数据输入变量的预设取值范围；或提示并接受用户输入缺少数据输入变量的预设取值范围。

在对应的预设取值范围内随机生成预定数量个值对缺少数据的输入变量进行预定数量次赋值；其中，预定数量由用户输入或预先设置。

在一个技术方案中，取值范围获取模块 3 还用于从样本数据库中随机选取预定数量个训练输入数据，然后提取其中该输入变量对应的数据对缺少数据输入变量进行预定数量次赋值；其中，预定数量由用户输入或将小于等于样本数据库中输入数据的数量的 $P\%$ 作为预定数量，P 小于等于 100，样本数据库中的输入数据都是完整的输入数据。

(4) 数据输入模块 4，用于将多组完整输入数据输入到预先训练好的深度学习模型中，计算得到多组输出数据。

在一个技术方案中，数据输入模块 4 还用于统计多组输出数据中各类输出数据所占的比例；将比例最高的一类输出数据作为预测结果；将预测结果及各类输出数据及其相应比例输出给用户。

(5) 在一个技术方案中，上述深度学习模型装置还包括训练模块 5，用于对深度学习模型进行初始化，获取训练数据，通过训练数据对初始化后的深度学习模型进行训练得到预先训练好的深度学习模型。



第4章

深度学习模型的预报预测大数据应用

深度学习模型的应用很多，其中深度学习模型的预报预测大数据应用有着非常重要的应用价值和前景。本章给出了两种深度学习模型的预报预测大数据应用：基于大数据深度学习和分析场的气象预报、分析预测文化冲突的深度学习。基于大数据深度学习和分析场的气象预报方法包括：获取预报区域最近设定时长内的气象分析场序列；将最近气象分析场序列输入预报区域的深度神经网络进行深度学习，得到预报区域的深度神经网络的输出；将输出作为预报区域自当前时刻开始的预报时长内的预报场序列。该系统包括：分析场获取模块，用于获取预报区域最近设定时长内的气象分析场序列；预报模块，用于将最近气象分析场序列输入预报区域的深度神经网络进行深度学习，得到预报区域的深度神经网络的输出；将输出作为预报区域自当前时刻开始的预报时长内的预报场序列。利用上述方法和系统，气象预报更加准确。分析预测文化冲突的深度学习方法包括：获取待预测用户的文化差异值，文化差异值反映了待预测用户迁入地和迁出地之间的文化差异；获取待预测用户的文化素养值、同类群体大小值以及过渡时长值；将文化差异值、文化素养值、同类群体大小值以及过渡时长值输入到预先训练好的深度学习模型中，输出待预测用户的文化冲突事件类型数据。通过对深度学习模型的运用，能够在获取到个人的某些特定数据的情况下，对个人进行移民时可能发生的文化冲突事件类型进行预测，进而为预防和消解文化冲突提供决策辅助。

4.1　基于大数据深度学习和分析场的气象预报

天气变化与人们的生产活动、社会活动、军事活动以及日常生活都有着十分密切的关系。一直以来，人们总是想方设法去预测未来的天气变化，以利用有利的天气，同时提前预

防不利天气带来的气象灾害。

传统技术中，人们预测天气采用数值预测的方法，数值预测的原理为：通过天气图或者卫星云图获取气象图像，并在一定的初值和边值条件下，利用大型计算机做数值计算，求解描述天气演变过程的流体力学和热力学的方程组，预测未来一定时段的大气运动状态和天气现象的方法。上述的数值预测方法是基于数学模型进行模式推导的，误差较大[15,16]。

本技术研究并给出了一种新的气象预报方法和系统。具体方法包括：获取预报区域的最近设定时长内的气象分析场序列；将最近气象分析场序列输入预报区域的深度神经网络进行深度学习，得到预报区域的深度神经网络的输出；将输出作为预报区域的自当前时刻开始的预报时长内的预报场序列。具体系统包括：分析场获取模块，用于获取预报区域的最近设定时长内的气象分析场序列；预报模块，用于将最近气象分析场序列输入预报区域的深度神经网络进行深度学习，得到预报区域的深度神经网络的输出；将输出作为预报区域的自当前时刻开始的预报时长内的预报场序列。利用上述方法和系统，气象预报更加准确。

4.1.1　基于大数据深度学习和分析场的气象预报方法

1. 基于大数据深度学习和分析场的气象预报的基本方法

如图 4-1 所示为本技术方案的气象预报方法流程图。该方法包括：

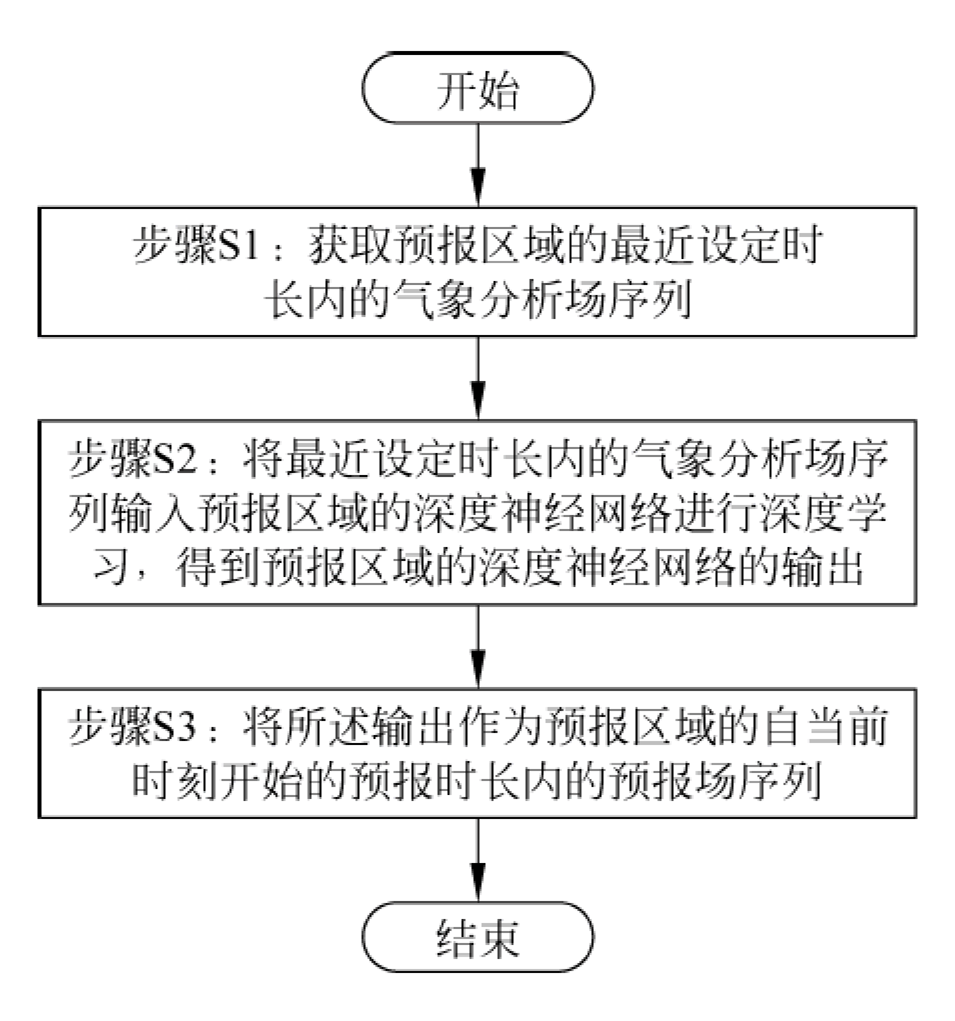


图 4-1　气象预报方法流程图

（1）步骤 S1，获取预报区域的最近设定时长内的气象分析场序列。预报区域指准备作出气象预报的区域，范围可根据需求确定，例如中国可按照省、市等行政区域进行划分。最近设定时长内是指以当前时间为分界点，在当前时间之前的最近的一个设定时长的时间段内。气象分析场指的是用实况资料作为分析数据，是真实的、确定的。气象分析场在气象预报领域的含义是所属领域的技术人员所熟知的。

气象分析场序列由多个时间点的气象分析场有序组成。例如，设定时长是 60min，且假定气象分析场的时间分辨率是 1min，那么 60min 内会有 60 个气象分析场，这 60 个气象分

析场按照时间先后顺序排列就组成了一个时长为 60min 的气象分析场序列。预报区域例如深圳市,预报时长例如 12h,即从当前时刻开始的将来 12h。

参考图 4-2,假设当前时刻是 4 月 11 日 12 点 0 分,预报区域的最近设定时长内气象分析场序列就是 4 月 11 日 11 点 0 分至 4 月 11 日 12 点 0 分的气象分析场序列。

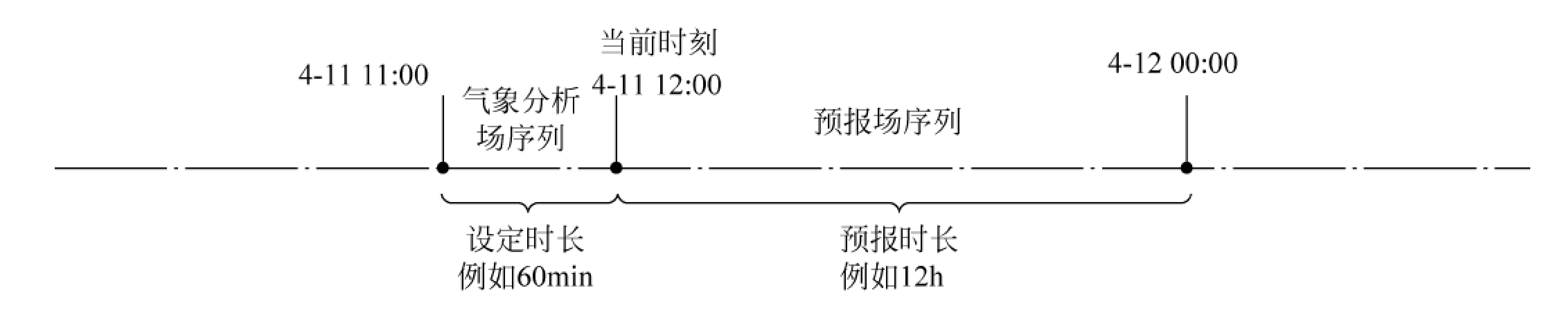


图 4-2 各时间点和时间段在时间轴上的相对位置示意图

(2) 步骤 S2,将最近设定时长内的气象分析场序列输入预报区域的深度神经网络进行深度学习,得到预报区域的深度神经网络的输出。深度神经网络特指预报区域的深度神经网络,其以该预报区域的气象分析场序列为输入,根据输入得到深度神经网络的输出。

(3) 步骤 S3,将输出作为预报区域的自当前时刻开始的预报时长内(例如,预报时长为 12h 时,则为 4 月 11 日 12 点 0 分—4 月 12 日 0 点 0 分)的预报场序列。

采用深度学习和深度神经网络对气象进行预测,由于深度学习和深度神经网络的预测精度非常高(这已经在很多应用中得到了印证,例如图像识别,而气象分析场序列与图像数据没有本质上的区别。例如将分析场中每个点作为一个像素点,用该像素点的 RGB 表示分析场中该点的属性值,则每个分析场都可以用一个图像来表示。而气象分析场序列是由多个分析场有序组成,等同于多个图像有序组成一个更大的图像),所以可以高精度地预测气象,这种预报结果可以与模式预报的结果一起作为预报的参考依据,使得预报员能做出更为准确的天气预报。

2. 基于大数据深度学习和分析场的气象预报的详细方法

如图 4-3 所示,步骤 S1 包括:

(1) 步骤 S1-1,获取预报区域。预报区域可以由用户指定。确定预报区域的意义还在于选择对应的深度神经网络和输入相应的气象分析场序列。

(2) 步骤 S1-2,确定预报时长。预报时长可以由用户指定。预报时长可以包括未来 12h、24h、48h 以及 72h 等。

(3) 步骤 S1-3,判断气象分析场是否存在,若气象分析场存在,则执行步骤 S1-4;否则执行步骤 S1-5。

(4) 步骤 S1-4,直接获取气象分析场。

(5) 步骤 S1-5,将气象分析场所在时刻的在分析场作为分析场。在气象分析场是用实况资料进行插值反演等,因为实况资料的站点分布不均匀,在气象分析场是经过插值后均匀分布的。气象分析场在气象预报领域的含义也是所属领域的技术人员所熟知的。

如图 4-4 所示,对预报区域的深度神经网络进行训练。深度神经网络的特点是:经过大量有效的数据的训练后,其预测的准确性会更高。在执行上述步骤 S2 之前,需要保证预报区域的深度神经网络具有相当程度的预测准确率,否则应该首先执行大量的训练过程,再将深度神经网络用于气象预报。

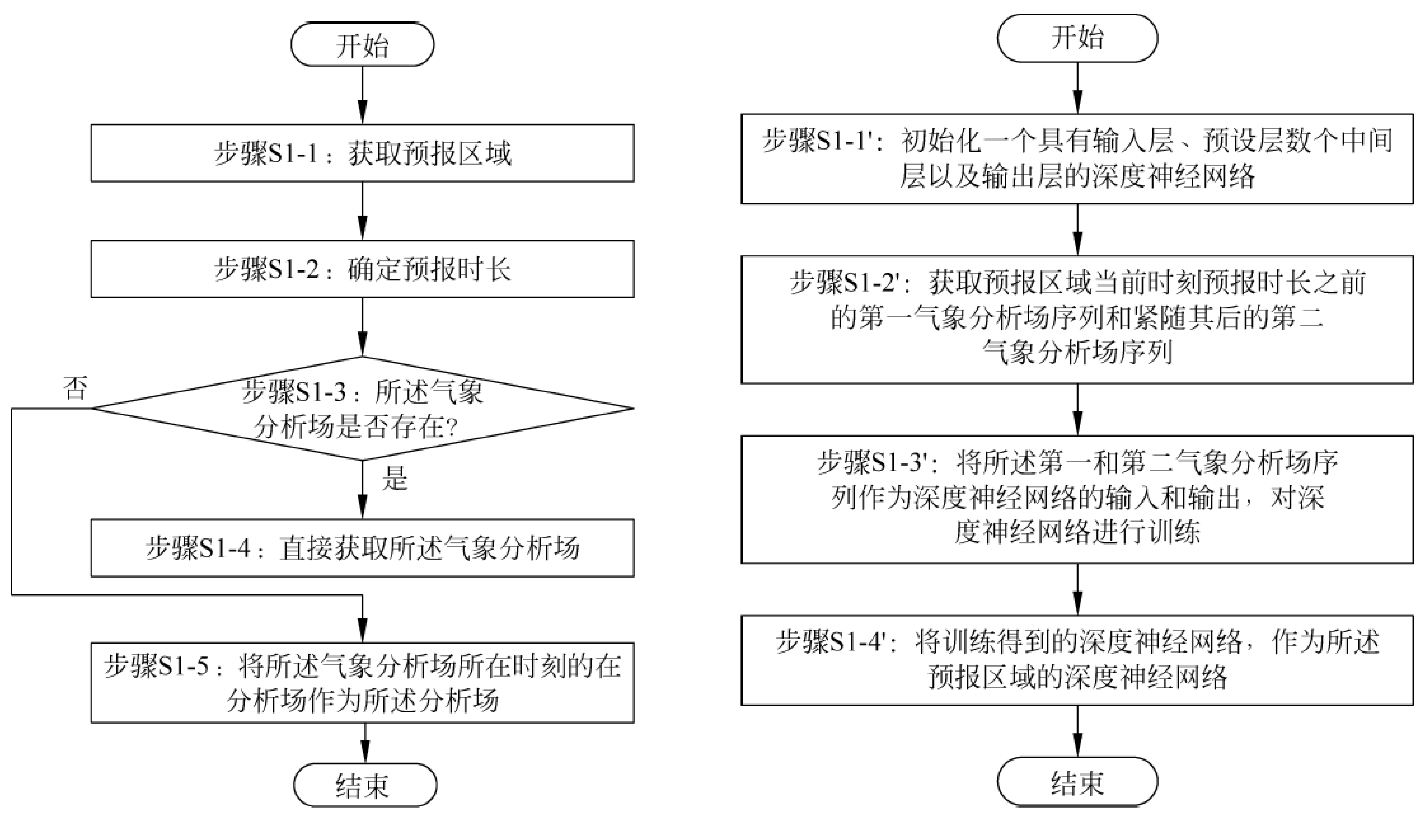


图 4-3　步骤 S1 的具体流程图　　　　图 4-4　训练深度神经网络的流程图

步骤 S1'包括：

子步骤 S1-1'：初始化一个具有输入层、预设层数中间层以及输出层的深度神经网络。预设层数小于设定时长内气象分析场序列中气象分析场的数量。

子步骤 S1-2'：获取预报区域当前时刻预报时长之前的第一气象分析场序列和紧随之后的第二气象分析场序列。其中，第一气象分析场序列为任一符合预设条件的设定时长内的气象分析场序列。预设条件为无条件或季节条件。季节条件是指，当前预报时间所在的季节与输入的气象分析场序列数据应该在季节上一致，例如 100 天之内。

参考图 4-5，例如当前时刻是 4 月 11 日 12 点 0 分，预报时长为 12h，则 12h 之前是 4 月 11 日 0 点 0 分之前。那么在这个例子中，当时时刻预报时长之前，就是 4 月 11 日 0 点 0 分之前。在 4 月 11 日 0 点 0 分之前，取任一设定时长内(例如 4 月 9 日 23 点 0 分—4 月 10 日 0 点 0 分)的气象分析场序列作为第一气象分析场序列，其之后的预报时长内(例如 4 月 10 日 0 点 0 分—4 月 10 日 12 点 0 分)的气象分析场序列为第二分析场序列。

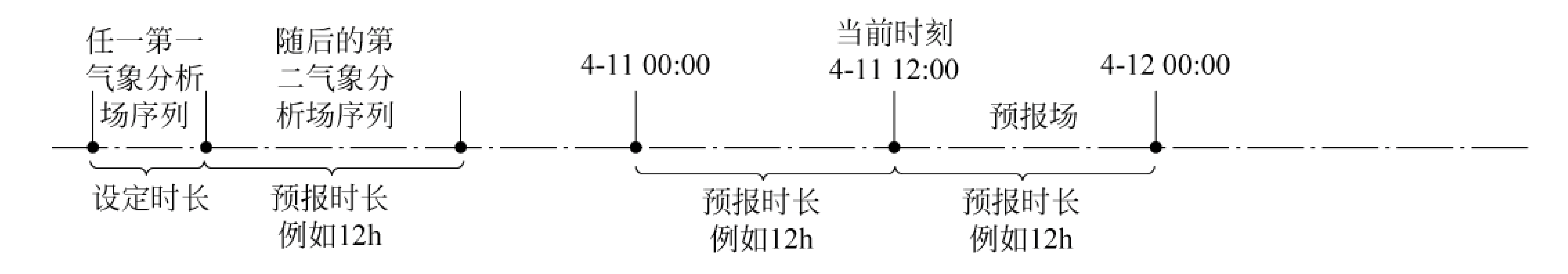


图 4-5　训练过程的各时间点和时间段在时间轴上的相对位置示意图

子步骤 S1-3′：将第一和第二气象分析场序列作为深度神经网络的输入和输出，对深度神经网络进行训练；获取多个第一气象分析场序列和对应的第二气象分析场序列，根据上述训练方法对深度神经网络进行重复训练。

具体而言，在将第一气象分析场序列输入到输入层、第二气象分析场序列输入到输出层之后，进行以下操作：

首先，在第一气象分析场序列中选取预设层数数量的气象分析场。在子步骤 S1-1′中，初始化一个具有预设层数中间层的深度神经网络。本步骤中，从第一气象分析场序列中选取该预设层数数量的气象分析场，获取预设层数个气象分析场对应的时间点，将时间点中离当前时刻最远的第 k 个时间点之后预报时长内的气象分析场序列输入第一深度神经网络的第 k 个中间层，其中，k 是从 1 到预设层数的自然数。

为控制深度神经网络完成训练，获取当前时刻(例如 4 月 11 日 12 点 0 分)预报时长之前(例如，预报时长为 12h，12h 之前是 4 月 11 日 0 点 0 分之前)最近的符合预设条件的设定时长(例如 4 月 10 日 23 点 0 分—4 月 11 日 0 点 0 分)内的第三气象分析场序列及紧随的预报时长内(例如 4 月 11 日 0 点 0 分—4 月 11 日 12 点 0 分)的第四气象分析场序列。

深度神经网络完成训练的条件为：

将第三气象分析场序列输入深度神经网络获得输出结果与对应的第四气象分析场序列之间的对比误差不大于停止阈值。深度神经网络在进行训练的同时，也可以用历史数据进行预测，并将预测的结果与已有的历史数据进行比较，如果预测结果与历史数据非常接近，则说明深度神经网络训练完成，可以将其用于预报。具体而言，用于训练的数据存储在历史气象大数据中。气象分析场序列均是由服务器中预存的气象分析场中查找获取的，其中，气象分析场是通过大数据进行存储和获取的，也就是采集任一时间任一预报区域的气象分析场存入大数据，从大数据中获取任一时间任一预报区域的气象分析场。

具体而言，建立气象分析场的 HBASE 大数据库表，并将采集到的气象分析场及其对应时间存入气象分析场的 HBASE 大数据库表，同时可以根据气象分析场的时间从气象分析场的 HBASE 大数据库表中获取对应时间的气象分析场。

子步骤 S1-4′：将训练得到的深度神经网络，作为预报区域的深度神经网络。

4.1.2 基于大数据深度学习和分析场的气象预报系统

如图 4-6 所示的为一技术方案的气象预报系统模块图，该气象预报系统包括分析场获取模块 1、预报模块 2 和深度神经网络 3。

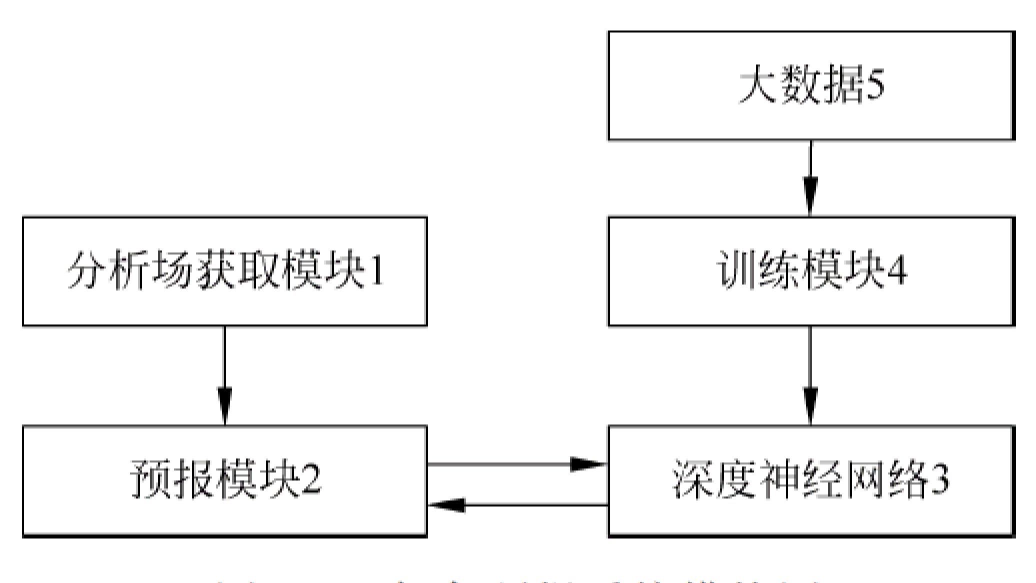


图 4-6 气象预报系统模块图

(1) 分析场获取模块 1 用于获取预报区域最近设定时长内的气象分析场序列。预报区域指准备作出气象预报的区域，范围可根据需求确定。例如，中国可按照省、市等行政区域

进行划分。最近设定时长内是指以当前时间为分界点，在当前时间之前最近的一个设定时长的时间段内。气象分析场指的是用实况资料进行的分析数据，是真实的、确定的。气象分析场在气象预报领域的含义是所属领域的技术人员所熟知的。

如图 4-7 所示，分析场获取模块 1 包括预报区域选择单元 1-1、预报时长确定单元 1-2 以及分析场生成单元 1-3。

预报区域选择单元 1-1 用于获取预报区域。确定预报区域的意义还在于选择对应的深度神经网络和输入相应的气象分析场序列。

预报时长确定单元 1-2 用于确定预报时长。预报时长可以由用户指定。预报时长可以是未来 24h、48h 以及 72h 等。

分析场生成单元 1-3 用于若气象分析场存在，则直接获取气象分析场，否则将气象分析场所在时刻的在分析场作为分析场。在分析场是用实况资料进行插值反演等，因为实况资料的站点是分布不均匀的，在分析场是进过插值后是均匀分布的。在分析场在气象预报领域的含义也是所属领域的技术人员所熟知的。

（2）预报模块 2 用于将最近气象分析场序列输入预报区域的深度神经网络进行深度学习，得到预报区域深度神经网络的输出。深度神经网络特指预报区域的深度神经网络，其以该预报区域的气象分析场序列为输入，根据输入得到深度神经网络的输出。

（3）该气象预报系统还可以包括对预报区域的深度神经网络 3 进行训练的训练模块 4。如图 4-8 所示，训练模块 4 包括初始化单元 4-1、气象分析场获取单元 4-2 和训练控制单元 4-3。深度神经网络的特点是，经过大量有效的数据训练后，其预测的准确性会更高。在进行气象预报前，需要保证预报区域的深度神经网络具有相当程度的预测准确率，否则应该首先执行大量的训练过程，再将深度神经网络用于气象预报。

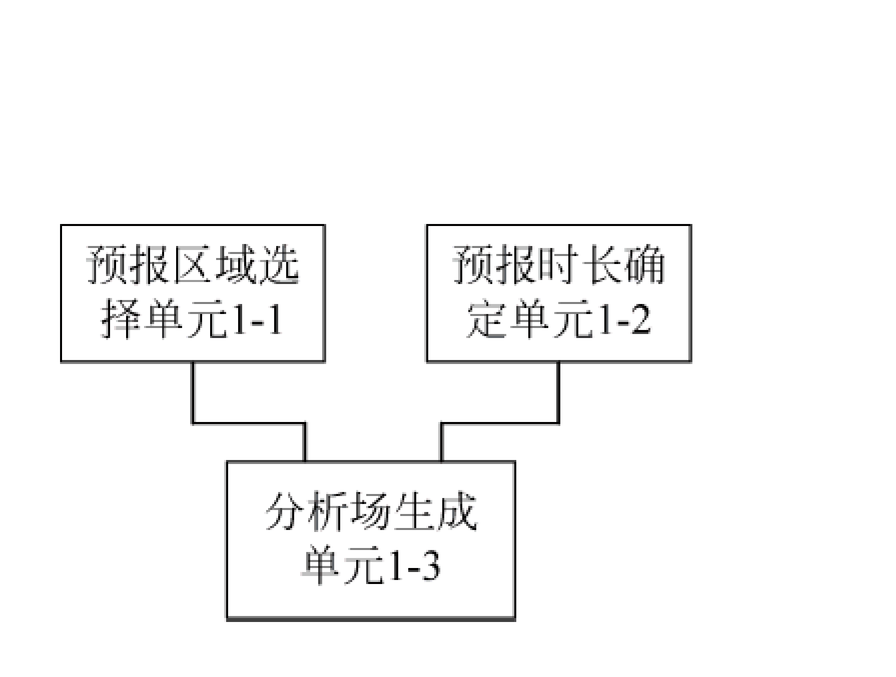


图 4-7　分析场获取模块的模块图

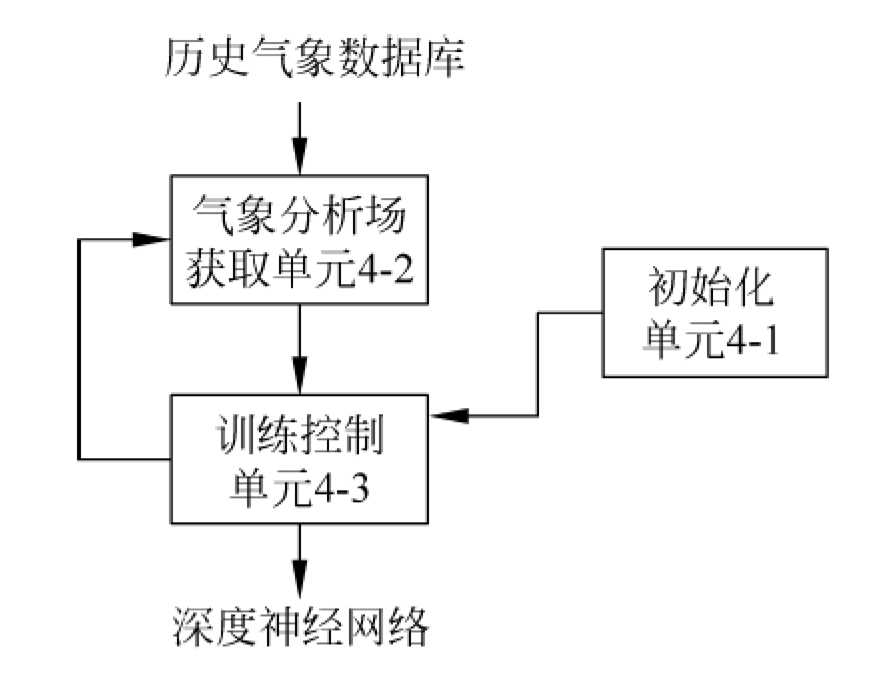


图 4-8　训练模块的模块图

初始化单元 4-1 用于初始化一个具有输入层、预设层数个中间层以及输出层的深度神经网络；预设层数小于设定时长内的气象分析场序列中气象分析场的数量。

气象分析场获取单元 4-2 用于获取预报区域当前时刻预报时长之前的第一气象分析场序列和紧随之后的第二气象分析场序列，其中第一气象分析场序列为任一符合预设条件的设定时长内的气象分析场序列。

训练控制单元 4-3 用于将第一和第二气象分析场序列作为深度神经网络的输入和输出，对深度神经网络进行训练；并获取多个第一气象分析场序列和对应的第二气象分析场序列，对深度神经网络进行重复训练；将训练完成的深度神经网络，作为预报区域的深度神

经网络。

训练控制单元还获取当前时刻预报时长之前最近的符合预设条件的设定时长内的第三气象分析场序列及紧随之后的第四气象分析场序列。深度神经网络完成训练的条件为：第三气象分析场序列输入深度神经网络所得的输出结果与对应的第四气象分析场序列之间的对比误差不大于停止阈值。

（4）该气象预报系统还可以包括历史气象大数据5，用于训练的数据存储在历史气象大数据5中。采用大数据5，可以利用大数据进行气象预报，提高预测的准确性。

上述气象预报方法和系统，采用深度神经网络对最近的预报时长内的气象分析场序列进行深度学习，得到从当前时间开始的预报时长内的预报场。由于深度学习和深度神经网络的预测精度非常高（这已经在很多应用中得到了印证，例如图像识别，而气象分析场序列与图像数据没有本质上的区别。例如将气象分析场中每个点作为一个像素点，用该像素点的RGB表示气象分析场中该点的属性值，则每个气象分析场都可以用一个图像来表示。而气象分析场序列是由多个气象分析场有序组成的，等同于多个图像有序组成一个更大的图像），所以可以高精度地预测气象，预测结果更加准确。这种预测结果可以与模式预报的结果一起作为预报员预报的参考依据，使得预报员能做出更为准确的天气预报。

4.2 基于深度学习模型的文化冲突分析预测

人工神经网络（Artificial Neural Networks，ANN）也简称为神经网络（NN）或连接深度学习模型（Connection Model），它是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学深度学习模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到信息处理的目的。人工神经网络在工程与学术界也常直接简称为“神经网络”或类神经网络，可以用于分析预测。

传统技术中对于文化认同、文化融合以及文化冲突[17,18]等都是定性分析，没有定量分析，没有形成对个人影响的判断。而很多社会问题往往都是文化冲突作用于个人所形成的，因此无法分析文化冲突对个人的影响，从而无法定量地起到辅助决策和预防冲突的作用。

本技术研究并给出了一种新的分析预测文化冲突的深度学习方法、系统。方法包括：获取待预测用户的文化差异值，文化差异值反映了待预测用户迁入地和迁出地之间的文化差异；获取待预测用户的文化素养值、同类群体大小值以及过渡时长值；将文化差异值、文化素养值、同类群体大小值以及过渡时长值输入预先训练好的深度学习模型中，输出待预测用户的文化冲突事件类型数据。通过对深度学习模型的运用，能够在获取个人某些特定数据的情况下，对个人进行移民时可能发生的文化冲突事件类型进行预测，进而为预防和消解文化冲突提供决策辅助。

4.2.1　基于深度学习模型的文化冲突分析预测方法

1. 分析预测文化冲突的深度学习模型的基本方法

如图 4-9 所示，在一个技术方案中，提供了一种分析预测文化冲突的深度学习方法，包括：

（1）步骤 S1，获取待预测用户的文化差异值，文化差异值反映了待预测用户迁入地和迁出地之间的文化差异。

在数据库中存储有多组迁入地与迁出地之间的文化差异值。当获知待预测用户的迁入地和迁出地时，则可在数据库中查找到跟待预测用户的迁入地与迁出地一致的那组文化差异数据，从而获得待预测用户的文化差异值，此文化差异值反映了待预测用户迁入地和迁出地之间的文化差异。

（2）步骤 S2，获取待预测用户的文化素养值、同类群体大小值以及过渡时长值。

步骤S1：获取待预测用户的文化差异值。文化差异值反映了待预测用户迁入地和迁出地之间的文化差异

↓

步骤S2：获取待预测用户的文化素养值、同类群体大小值以及过渡时长值

↓

步骤S3：将文化差异值、文化素养值、同类群体大小值以及过渡时长值输入预先训练好的深度学习神经网络中，输出待预测用户的文化冲突事件类型数据

图 4-9　分析预测文化冲突方法的流程示意图

（3）步骤 S3，将文化差异值、文化素养值、同类群体大小值以及过渡时长值输入预先训练好的深度学习模型中，输出待预测用户的文化冲突事件类型数据。

对于预先训练好的深度学习模型，已经将文化差异、文化素养、同类群体以及过渡时长设置为深度学习模型的输入变量，将文化冲突事件类型设置为深度学习模型的输出变量。因此将获取的待预测用户的文化素养值、同类群体大小值以及过渡时长值的数据输入预先训练好的深度学习模型中时，预先训练好的深度学习模型会计算并输出对应的待预测用户的文化冲突事件类型数据。文化差异值反映了迁入地与迁出地的文化差异程度，文化素养反映了该用户的文化程度，同类群体反映的是与该用户移民状态一致或类似的人群，过渡过长反映了该用户从迁出地的文化过渡到迁入地文化所需的时长。

2. 基于深度学习模型的文化冲突分析预测的详细方法

在一个技术方案中，在获取待预测用户的文化差异值之前，须获取数据库中的用户样本数据；根据用户样本数据获取用户的文化差异值、文化素养值、同类群体大小值、过渡时长值以及文化冲突事件类型数据对深度学习模型进行训练。

在数据库中存储有多个用户样本数据，用户样本数据中包含了多个用户的文化差异值、文化素养值、同类群体大小值、过渡时长值以及文化冲突事件类型数据。在深度学习模型中，预先对深度学习模型的输入变量与输出变量进行了配置，将文化差异、文化素养、同类群体以及过渡时长设置为深度学习模型的输入变量，将文化冲突事件类型设置为深度学习模型的输出变量。因此将多个用户的文化差异值、文化素养值、同类群体大小值、过渡时长值以及文化冲突事件类型数据分别作为深度学习模型的输入和预期输出，深度学习模型会根据多组数据进行训练。训练后的深度学习模型才能根据输入数据预测输出数据。对深度学习模型进行训练能使得获取的预测结果更为准确。

在一个技术方案中，根据用户样本数据获取用户的文化差异值、文化素养值、同类群体大小值、过渡时长值以及文化冲突事件类型数据对深度学习模型进行训练，包括：

根据用户样本数据获取用户的文化差异值、文化素养值、同类群体大小值以及过渡时长值，将其输入到深度学习模型中进行无监督训练。

从用户样本数据中获取与用户的文化差异值、文化素养值、同类群体大小值以及过渡时长值对应的文化冲突事件类型数据，将用户的文化差异值、文化素养值、同类群体大小值以及过渡时长值作为深度学习模型的输入数据，将文化冲突事件类型数据作为深度学习模型的预期输出，对深度学习模型进行有监督训练。

在无监督训练后，再对深度学习模型进行有监督训练。进行有监督训练时，会完整地提供深度学习模型的输入数据与预期输出。

用户样本数据中存储有多组用户数据，每组用户数据中的数据类型包括每个用户的文化差异值、文化素养值、同类群体大小值、过渡时长值以及文化冲突事件类型数据。但是，并不是每组用户数据都是完整的，可能存在有的用户数据中缺少文化冲突事件类型数据，这部分缺少文化冲突事件类型数据的用户数据可以用于对深度学习模型进行无监督训练，避免了数据的浪费。由于无监督训练预先训练了深度学习模型的特征提取能力，在进行了无监督训练后再进行有监督训练也能提升深度学习模型的训练效果，提升了训练后的深度学习模型预测输出数据的准确度。

在一个技术方案中，在获取待预测用户的文化差异值之前，须获取迁入地和迁出地的本地居民的问卷调查数据。问卷调查数据包括对预定选项的选择结果。根据如下公式计算待预测用户的文化差异值：

文化差异值＝选项数量差值/(参与问卷调查的本地居民人数×1/2×问卷调查选项的总和)

其中，选项数量差值为调查问卷上迁入地与迁出地本地居民之间对每一个选项的选择数量的差值绝对值相加和；问卷调查选项的总和为调查问卷中选项项数的总和。

文化差异值反映了待预测用户的迁入地和迁出地之间的文化差异。获取文化差异值的方法主要是通过问卷调查的方式，问卷调查可在线上或线下进行。问卷调查选取的调查对象主要是本地居民，本地居民指在本地出生长大，并到填写问卷调查之前还未移居至其他城市。本地居民的限定可根据研究人员与设计人员的考虑而定，例如研究人员限定从祖父辈开始定居在此城市的居民才算本地居民。

问卷调查上有多个预定的选择题与对应的多个选项。当本地居民填写完毕后，统计调查结果。具体的统计方式为，统计每个选项的选择人数并与城市对应统计，例如第一题有2个选项，选项A和选项B，统计选择选项A、选择选项B的北京居民人数，以及选择选项A、选择选项B的广州居民人数。再将每个选项迁入地居民与迁出地居民选择数量的差值的绝对值相加得到选项数量差值。例如有两道题，每道题有2个选项A和B，3个北京本地居民与3个广州居民参与了调查。第1题有1个北京本地人选A，2个北京本地人选B，2个广州本地人选A，1个广州本地人选B；第2题有3个北京本地人选A，3个广州本地人选B。那么，第1题的选项A的差值为2－1＝1，第1题的选项B的差值为2－1＝1；第2题的选项A的差值为3－0＝3，第2题的选项B的差值也为3－0＝3。将差值的绝对值全部相加得到选项数量差值，则选项数量差值为1＋1＋3＋3＝8。

再统计问卷调查选项的总和，即为问卷调查上全部题目的全部选项之和。例如，根据上述举例，那么调查选项的总和就为 2+2=4。那么北京与广州之间的文化差异值为 8 除以 3 再除以 4，结果为 2/3，即北京与广州之间的文化差异为 2/3。

通过获取问卷调查的数据，并通过计算机系统进行自动计算来统计得到多个迁入地与迁出地之间的文化差异值，在对待预测用户的文化冲突事件类型进行预测时，则可直接使用已经统计得到的文化差异值，减少了每一次都需要计算文化差异值的步骤，使得预测操作更为简单。

在一个技术方案中，文化素养值为对待预测用户的学历进行量化得到的值；同类群体大小值为与待预测用户的迁入地与迁出地一致的用户的数量。

待预测用户的文化素养值，是指根据待预测用户的学历量化得到的值。量化标准可根据研究者与设计者的考虑而定。例如，将小学学历量化值为 0，初中学历量化值为 1，高中学历量化值为 2，大学学历量化值为 3，硕士学历量化值为 4，博士学历量化值为 5，博士后量化为 6。因此可根据指定的学历量化标准，在获知待预测用户的学历信息后，可将学历用对应的数值替代。学历代表了一个人接受教育的程度，也能在一定程度上代表一个人的文化素养。

待预测用户的同类群体大小值，是指与待预测用户的迁出地与迁入地均一致的其他用户的数量。例如，待预测用户 A 的迁出地为北京，迁入地为西安，那么待预测用户 A 的同类群体大小值就是迁出地为北京，迁入地为西安的其他用户数量，也可以称之为迁出地与迁入地一致的移民数量。具体是指迁出地与迁入地的城市一致的移民数量。例如，当待预测用户 A 的迁出地是北京，而迁入地是广州，那么只要是迁出地属于北京市的范围区域，迁入地属于广州市的范围区域的移民都属于待预测用户 A 的同类群体，而不需要考虑这些移民具体的迁出地所在的小区之类的。同类群体越大，说明从某一个地方集体移民过来的可能性也较大，这种情况下，这部分人在较长时间还是保持着自己过去的文化的可能性也越大，那么融入迁入地文化的难度也越大。故有必要对同类群体的大小进行统计与分析。这种采用较容易获取到的数据对待预测用户的文化冲突进行分析的方法，使得调查结果会更客观也更准确。

在一个技术方案中，过渡时长值根据如下公式计算得到：

$$\text{过渡时长值}=T_1\times k\%+T_2\times(1-k\%)$$

其中，T_1 为待预测用户在移民到迁入地之前在迁入地停留的时间天数；T_2 为待预测用户在移民到迁入地之前接受过迁入地文化相关的学习时间天数；$k\%$为大于 0 小于 1 的预设权值。

过渡时长值，是为了获知待预测用户在移民至迁入地之前，对迁入地的文化、环境等了解的程度。$k\%$是区分 T_1 与 T_2 之间的权重，若是研究人员或设计人员认为 T_1 与 T_2 的权重一样，那么可将 k 设为 50。获知到待预测用户在移民至迁入地之前停留的时间 T_1 与接受迁入地文化相关的学习时间 T_2，那么此待预测用户的过渡时长值也可获取到。T_1 与 T_2 的值越大，过渡时长值也会越大，表明此待预测用户对于迁入地的了解较多。过渡时长值也能为文化冲突的预测提供一种新的思路，当用户移民至某个迁入地时，是否发生文化冲突与过渡时长存在着某种关联，也可为有关部门提供更准确的分析数据和更优的文化冲突解决方案。

3. 获取推荐过渡时长的方法

在一个技术方案中，图 4-10 给出了一种获取推荐过渡时长方法的流程示意图，该方法包括：

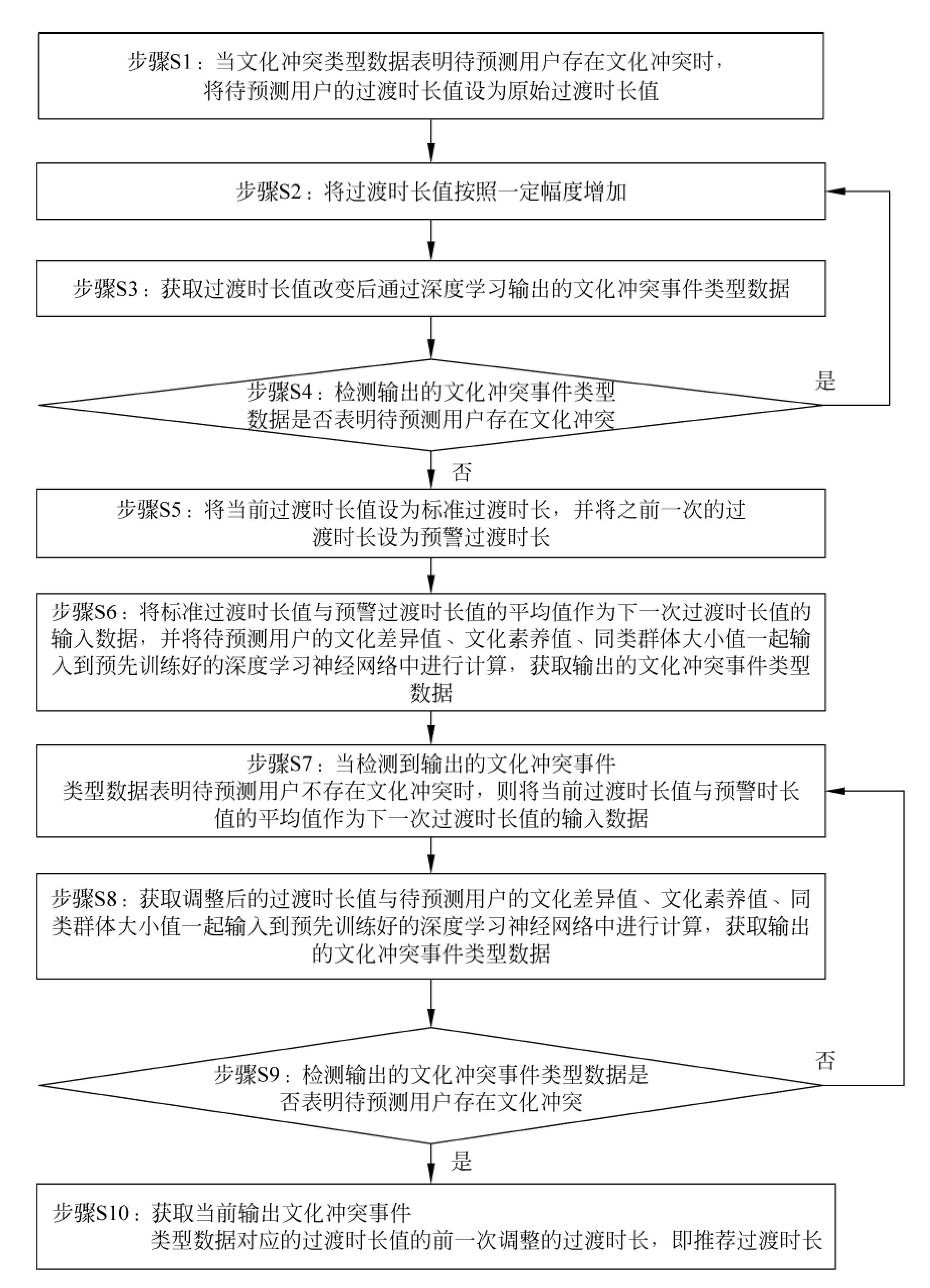


图 4-10 获取推荐过渡时长方法的流程示意图

（1）步骤 S1，当文化冲突事件类型数据表明待预测用户存在文化冲突时，将待预测用户的过渡时长值设为原始过渡时长值。

（2）步骤 S2，将过渡时长值按照一定幅度增加。

（3）步骤 S3，获取过渡时长值改变后通过深度学习输出的文化冲突事件类型数据。

(4) 步骤 S4,检测输出的文化冲突事件类型数据是否表明待预测用户存在文化冲突,若否,则执行步骤 S5;若是,则执行步骤 S2。

(5) 步骤 S5,将当前的过渡时长值设为标准过渡时长,并将之前一次的过渡时长设为预警过渡时长。

将待预测用户的数据对应输入到预先训练好的深度学习模型中后,获取对应输出的文化冲突事件类型数据。当输出的文化冲突事件类型数据表明该用户存在文化冲突时,则说明此用户还需要延长过渡时长,即学习迁入地文化的时间。先将待预测用户的过渡时长值进行按照一定数值固定增加。例如,待预测用户的过渡时长值为 1,设定增加幅度为 10,则将过渡时长值 1 每次增加 10,并将对应增加的数值 11、21、31、41 等作为过渡时长项的输入数据。同时,待预测用户的其他数据,即文化差异值、文化素养值、同类群体大小值等均保持不变,并与每次增加后的过渡时长值输入到预先训练好的深度学习模型中进行计算,获取输出的文化冲突事件类型数据,当输出的文化冲突事件类型数据表明该用户不存在文化冲突时,则停止过渡时长值的增加,并将当前过渡时长值设为标准过渡时长值,而前一次的值设为预警过渡时长值。

例如,设定的增加幅度为 10,当将过渡时长从 51 增加到 61 时,过渡时长值为 61 时输出的文化冲突事件类型数据表明该用户不存在文化冲突,那么就将 61 设为标准过渡时长值,而前一次的值,即 51,则设为预警过渡时长值。

(6) 步骤 S6,将标准过渡时长值与预警过渡时长值的平均值作为下一次过渡时长值的输入数据,并将待预测用户的文化差异值、文化素养值、同类群体大小值一起输入到预先训练好的深度学习模型中进行计算,获取输出的文化冲突事件类型数据。

(7) 步骤 S7,当检测到输出的文化冲突事件类型数据表明待预测用户不存在文化冲突时,则将当前过渡时长值与预警时长值的平均值作为下一次过渡时长值的输入数据。

(8) 步骤 S8,获取调整后的过渡时长值与待预测用户的文化差异值、文化素养值、同类群体大小值一起输入到预先训练好的深度学习模型中进行计算,获取输出的文化冲突事件类型数据。

(9) 步骤 S9,检测输出的文化冲突事件类型数据是否表明待预测用户存在文化冲突,若是,则执行步骤 S10;否则执行步骤 S7。

(10) 步骤 S10,获取当前输出文化冲突事件类型数据对应的过渡时长值的前一次调整的过渡时长,即推荐过渡时长。

得到标准过渡时长后,再将标准过渡时长值与预警过渡时长值的平均值作为下一次过渡时长值的输入数据,即采取二分法,得到下一次过渡时长值的输入数据。当输出的文化冲突事件类型数据仍然表明待预测用户不存在文化冲突时,例如输出的文化冲突事件类型数据均为 0 时,则继续采用二分法,将当前过渡时长值与预警过渡时长值的平均值作为下一次过渡时长值的输入数据。依次取值,直到输出的文化冲突事件类型数据表明待预测用户存在文化冲突,例如输出的文化冲突事件类型数据不均为 0 时,则获取上一次输出的文化冲突事件类型数据为 0 时对应的过渡时长值,将该过渡时长值设为推荐过渡时长值。而推荐过渡时长值与原始过渡时长值的差值即为该用户需要延期学习迁入地文化的时长。

通过改变过渡时长值研究当过渡时长值达到怎样的数值时,预测输出的该用户发生文

化冲突的可能性为0。首先将过渡时长值按照一定幅度增加，尽可能地增加过渡时长值，当输出的文化冲突事件类型数据均为0时，则表明这个过渡时长值对应的培训时间能让该用户发生文化冲突的可能为0。在此基础上采取二分法，缩小过渡时长值，从而选择最优的一个过渡时长，避免需要培训的过渡时间过长，也避免过渡时间不够不能达到减少文化冲突的效果。

在一个技术方案中，将用户的文化差异值比作电量，文化素养值比作能承受的功率，同类群体大小值比作电阻，过渡时长值比作该电量完全通过该电阻的时长，即电流时长，文化冲突事件类型比作电路故障类型。学历越高，文化素养越高，即能承受的功率也越高。当知道对应的电量与电流时长后，可计算出对应的功率，再将此功率与能承受的功率进行比较。当实际功率大于能承受的功率时，则会发生电路故障，即会发生文化冲突；当实际功率小于或等于能承受的功率时，则不会发生电路故障，即不会产生文化冲突。

例如，当移民A的数据如下：文化差异值为2，文化素养值为6，同类群体大小值为2500，过渡时长值为30，当移民A文化冲突事件类型数据为{1,1,1,0,0}时，对应地，电量为2，能承受的功率为6，电阻为2500，电流时长为30，电路故障类型为{1,1,1,0,0}。基于上述数据可计算得到电流＝电量/电流时长＝2/30，故实际功率＝电流^2×电阻＝2/30×2/30×2500＝100/9＝11，这种情况下，实际功率大于能承受的功率，则可能出现电路故障。这种采用物理函数类比的方式也能使得研究人员更好地了解这几项之间的关系，更有利于分析与调查。

在一个具体的技术方案中，根据将一个人身上的文化差异作为电量(如果这个差异需要消除，就需要这个电量通过这个人从正极流到负极，从而实现文化融合)，将这个人的文化素养作为能承受的功率(因为文化素养越高越能理解和承受住这种文化变化，大于这个功率就与熔丝一样，就会烧断，这个人就会出现问题，例如个人犯罪、个人心理问题、个人与社会的冲突；如果不超过这个功率，那么使电量通过这个人流动，这个人就能实现对新文化的认同或在这个人身上实现两种文化的融合)，将这个人的同类群体大小作为电阻(因为同类群体越大，被转变的难度就越大，例如从某一个地方集体移民来的人，这些人可能在较长时间还是保持着自己过去的文化)，该人从迁出地文化进入迁入地文化的过渡时长(例如从了解另一种文化到完全移民或生活到另一种文化所在的地区这中间所经历的时间)，作为该电量完全通过该电阻的时长，从而可以计算出电流，进而可以计算出功率。如果实际功率大于能承受的功率，就会发生电路故障，就对应着发生文化冲突；如果实际功率不大于能承受的功率，就不会发生电路故障，就对应着不发生文化冲突。虽然这些在意义上有对应关系，但因为单位不统一，难以归一化，因此需要深度学习模型进行处理。以电量、电阻、电流时长、可承受功率对应的实际量作为深度学习模型的输入，将是否发生文化冲突作为深度学习模型的输出(可以将输出分为几种类型，如犯罪、出现精神问题、与社会发生冲突等)，先通过已有样本进行学习训练，然后再在使用中对移民的文化冲突情况进行预测，进而可以对可能发生的文化冲突进行预防，并可以通过延长电流时间来消解文化冲突的发生。其意义在于能够对个人的文化冲突进行预测，进而为预防和消解文化冲突提供决策辅助。

4.2.2　基于深度学习模型的文化冲突分析预测系统

在一个技术方案中，如图 4-11 所示，提供了一种分析预测文化冲突的装置，包括：

(1) 文化差异值获取模块 1，用于获取待预测用户的文化差异值。文化差异值反映了待预测用户的迁入地和迁出地之间的文化差异。

(2) 待预测用户数据获取模块 2，用于获取待预测用户的文化素养值、同类群体大小值以及过渡时长值。

(3) 分析预测模块 3，用于将文化差异值、文化素养值、同类群体大小值以及过渡时长值输入到预先训练好的深度学习模型中，输出待预测用户的文化冲突事件类型数据。

在一个技术方案中，如图 4-12 所示，上述分析预测文化冲突的装置包括：

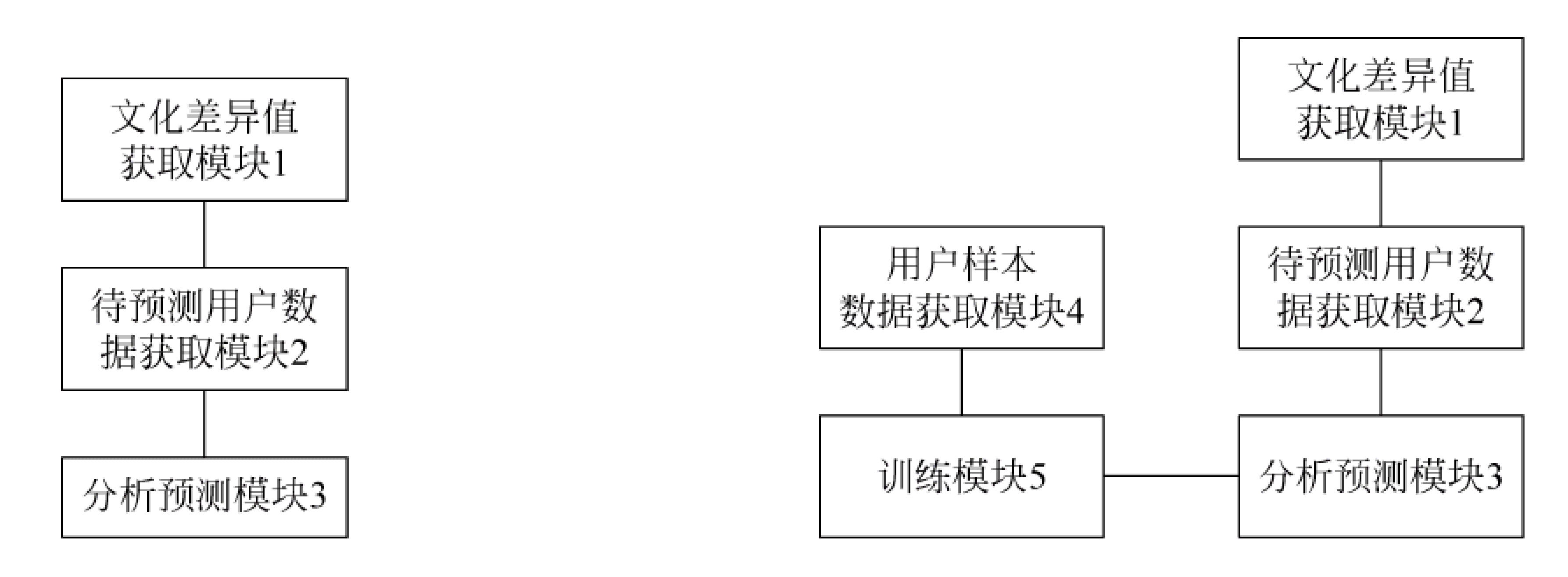


图 4-11　分析预测文化冲突装置的结构框图　　图 4-12　分析预测文化冲突装置的扩展结构框图

(4) 用户样本数据获取模块 4，用于获取数据库中的用户样本数据。

(5) 训练模块 5，用于根据用户样本数据获取用户的文化差异值、文化素养值、同类群体大小值、过渡时长值以及文化冲突事件类型数据对深度学习模型进行训练。

在一个技术方案中，上述训练模块 5 还用于根据用户样本数据获取用户的文化差异值、文化素养值、同类群体大小值以及过渡时长值，将其输入到深度学习模型中进行无监督训练；从用户样本数据中获取与用户的文化差异值、文化素养值、同类群体大小值以及过渡时长值对应的文化冲突事件类型数据，将文化差异值、文化素养值、同类群体大小值以及过渡时长值作为输入数据，将文化冲突事件类型数据作为深度学习模型的预期输出，对深度学习模型进行有监督训练。

(6) 推荐时长获取模块 6(可选模块，图 4-12 中未示出)，当文化冲突事件类型数据表明待预测用户存在文化冲突时，则将待预测用户的过渡时长值按照一定幅度增加；获取当文化冲突事件类型数据表明待预测用户不存在文化冲突时对应的过渡时长值为标准过渡时长；逐步调整标准过渡时长，并将每次调整后的标准过渡时长与用户的文化差异值、文化素养值、同类群体大小值一起作为输入数据输入到预先训练好的深度学习模型中，直到输出的文化冲突事件类型数据表明待预测用户存在文化冲突，此时获取前一次调整的标准过渡时长为推荐过渡时长。

在一个技术方案中，上述分析预测文化冲突的装置包括文化差异获取模块，用于获取迁入地和迁出地的本地居民的问卷调查数据。问卷调查数据包括对预定选项的选择结果。



第5章

深度学习模型的调度推荐大数据应用

深度学习模型的应用很多,其中深度学习模型的调度推荐大数据应用有着非常重要的应用价值和前景。本章给出了3种深度学习模型的调度推荐大数据应用:基于大数据和深度学习模型的云计算调度,基于大数据和深度学习模型的停车引导,基于大数据与深度学习的婚恋对象推荐。

5.1 基于大数据和深度学习模型的云计算调度

当前,云计算技术可谓是计算机服务领域最热门的话题之一。大到行业的领军企业,如IBM、Google,小到一些私人企业,甚至一些乐于追求新技术的技术人员,都在部署或研究云计算,希望通过云计算来巩固或提升自己在行业的地位。云计算环境中采用虚拟化技术,将服务器整体虚拟化为一个数据资源池,由于数据资源种类多、规模大,因此云计算数据资源调度成为云计算研究的热点之一。

进行云计算调度[19,20]时,最重要的是在调度之前预测一个云任务调度到某类云资源上去会占用多少云资源,以作为调度的依据。现有的云计算调度基本上都采用了预估式,通过预估某个云计算任务占用的云资源量(包括计算、内存、外存、网络等资源)和剩余云资源量进行调度。现有技术的这种预估方式一般是根据任务类型对占用的云资源量进行粗略估计的,但这种粗略估计并未结合过去对云资源的实际占用情况进行预测,因此往往是不准确的,导致基于此的调度也不一定能最大程度地利用云资源,调度效率不高。

如图5-1所示,本技术研究并给出了一种新的基于大数据和深度学习模型的云计算调度方法及系统,方法包括:获取需调度的云任务类型 T、需调度的 T 类型云任务的数量 M 以及拟调度到的云资源候选类型;从过去能力大数据库中检索出所有从类型 T 的云任务调

度到候选类型的云资源上的调度记录；根据检索出的所有过去调度记录，采用深度学习模型预测调度数量 M 的 T 类型云任务到候选类型的云资源时会占用候选类型云资源的数量；根据预测的结果完成云计算调度。本技术基于过去能力大数据和深度学习模型，通过过去不同类型的云任务调度到不同云资源上时对云资源的占用情况这一客观事实来进行预测，更加准确，调度效率更高。本技术可广泛应用于云计算领域。

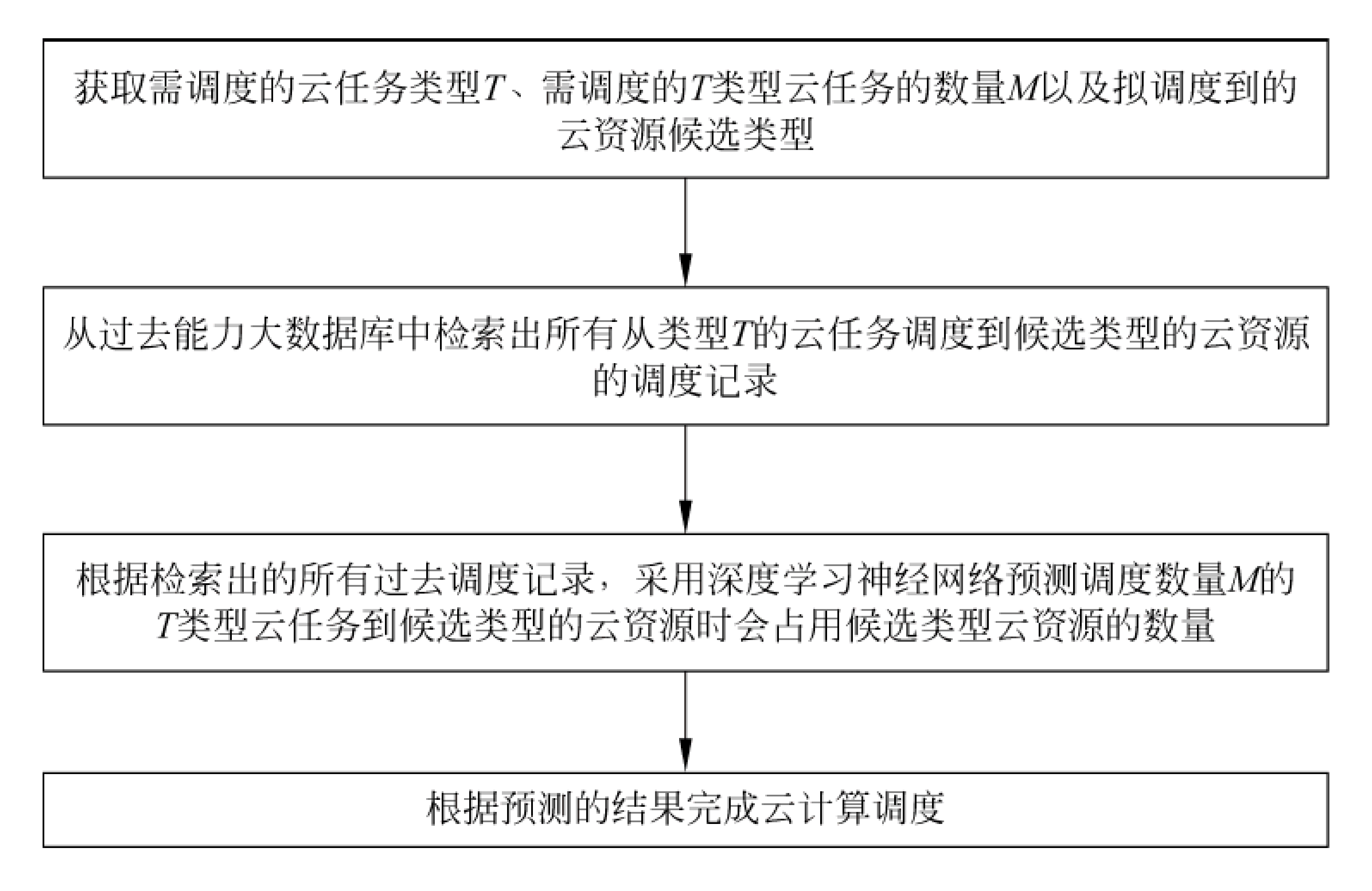


图 5-1　基于大数据和深度学习模型的云计算调度方法的整体流程图

5.1.1　基于大数据和深度学习模型的云计算调度方法

如图 5-2 所示，本技术云计算调度方案的一种具体技术方案的实现步骤如下：

（1）步骤 S1，将过去每次调度中不同类型的云任务调度到不同类型的云资源上时云任务的类型及数量、云资源的类型及占用数量存入过去能力大数据库。

其中，过去能力大数据库用于存储过去能力大数据。过去能力大数据，是指云调度的当前时间(即需进行云调度预测的时间)之前的云调度记录。云调度记录包括调度日期及时间、云任务类型、T 类型云任务的数量、云资源候选类型和占用候选类型云资源的数量等。例如，过去能力大数据包括：

日期及时间 1，云任务类型 T_{11}，T 类型云任务的数量 M_{11}，云资源候选类型 R_{11}，候选类型云资源的数量 N_{11}；

日期及时间 2，云任务类型 T_{21}，T 类型云任务的数量 M_{21}，云资源候选类型 R_{21}，候选类型云资源的数量 N_{21}；

……

（2）步骤 S2，获取需调度的云任务类型 T、需调度的 T 类型云任务的数量 M 及拟调度到的每一云资源候选类型。

本具体技术方案拟调度到的云资源类型至少一个，可以是一种类型或多种备选类型，云资源候选类型可用变量 R 表示。

步骤S1：将过去每次调度中不同类型的云任务调度到不同类型的云资源上时云任务的类型及数量、云资源的类型及占用数量存入过去能力大数据库

↓

步骤S2：获取需调度的云任务类型T、需调度的T类型云任务的数量M及拟调度到的每一云资源候选类型

↓

步骤S3：从过去能力大数据库中检索出所有从T类型的云任务调度到候选类型的云资源上的调度记录

↓

步骤S4：根据检索出来的每个调度记录中云任务的数量及云资源的占用情况，预测调度数量M的T类型云任务到候选类型云资源时会占用候选类型云资源的数量

↓

步骤S5：将预测出的每一候选类型云资源对应的数量作为调度相同数量的类型T云任务到该候选类型云资源时会占用该候选类型云资源的数量输出给云计算调度系统，云计算调度系统根据预测的结果从所有候选类型云资源中选择一个类型云资源进行调度

↓

步骤S6：在数量的T类型云任务被调度且运行完成后，获取数量M的类型T云任务实际占用的被调度的候选类型云资源的数量，并存入过去能力大数据库

图 5-2　具体技术方案的实现流程图

(3) 步骤 S3，从过去能力大数据库中检索出所有从 T 类型的云任务调度到候选类型的云资源上的调度记录。具体包括：

日期及时间 k_1，云任务类型 T，T 类型云任务的数量 M_1，云资源候选类型，候选类型云资源的数量 N_1；

日期及时间 k_2，云任务类型 T，T 类型云任务的数量 M_2，云资源候选类型，候选类型云资源的数量 N_2；

……

如果从 T 类型的云任务调度到候选类型的云资源上的所有调度记录的个数小于第一预设阈值，则检索出从与 T 类型匹配度大于符合第二预设条件的类型的云任务调度到与候选类型匹配度大于符合第三预设条件的类型的云资源上的所有调度记录，这样做的目的是确保有足够的数据训练预设深度学习模型。其中，与 T 类型匹配度大于符合第二预设条件的类型是指，如果 T 类型是“打开 OFFICE2007”，第二预设条件为任务名称匹配，那么类型“打开 OFFICE2003”属于与 T 类型匹配度符合第二预设条件的类型，因为二者的任务名称“打开 OFFICE”是匹配的，只是任务版本号不一致。与候选类型匹配度符合第三预设条件的类型是指，如果候选类型是“Windows 2003”，第三预设条件为资源名称匹配，那么类型“Windows 2008”属于与候选类型匹配度符合第三预设条件的类型，因为二者的资源名称

“Windows”是匹配的，只是资源版本号不一致。

（4）步骤S4，根据检索出来的每个调度记录中云任务的数量及云资源的占用情况，预测调度数量 M 的 T 类型云任务到候选类型云资源时会占用候选类型云资源的数量。

步骤S4-1，初始化预设深度学习模型，使得该预设深度学习模型的输入节点能接受 T 类型云任务的数量作为输入，并使得输入节点能接受候选类型云资源的数量作为输出。

步骤S4-2，将检索出来的每个调度记录按照时间先后顺序进行排序，并按照排序的先后依次作为训练数据对预设深度学习模型进行训练。

例如，将检索出来的每个调度记录按照时间先后顺序进行排序后得到：

日期及时间 k_1，云任务类型 T，T 类型云任务的数量 M_1，云资源候选类型，候选类型云资源的数量 N_1；

日期及时间 k_2，云任务类型 T，T 类型云任务的数量 M_2，云资源候选类型，候选类型云资源的数量 N_2；

……

训练时，首先进行如下无监督训练：

T 类型云任务的数量 M_1 作为训练输入数据，对预设深度学习模型进行无监督训练；

T 类型云任务的数量 M_2 作为训练输入数据，对预设深度学习模型进行无监督训练；

……

然后进行如下有监督训练：

T 类型云任务的数量 M_1 作为训练输入数据，候选类型云资源的数量 N_1 作为训练预期输出数据，对预设深度学习模型进行有监督训练；

T 类型云任务的数量 M_2 作为训练输入数据，候选类型云资源的数量 N_2 作为训练预期输出数据，对预设深度学习模型进行有监督训练；

……

因为越是最近的训练数据，对调度预测的参考价值越大，所以本具体技术方案需要按照时间先后顺序进行排序，并按照排序的先后依次作为训练数据对预设深度学习模型进行训练，这样做的好处是：越是最近时间的训练数据越是最后对深度学习模型进行训练，即令最近时间的训练数据对训练后的深度学习模型的影响更大，从而使得训练后的深度学习模型更能反映出最近的训练数据的特征。

步骤S4-3，将需调度的云任务数量 M 作为训练后的预设深度学习模型的输入，并将训练后的预设深度学习模型计算得到的输出作为预测出的云资源的占用数量。

（5）步骤S5，将预测出的每一候选类型云资源对应的数量作为调度相同数量的类型 T 云任务到该候选类型云资源时会占用该候选类型云资源的数量输出给云计算调度系统，云计算调度系统根据预测的结果从所有候选类型云资源中选择一个类型云资源进行调度。

例如，有 m 个候选类型 R，第1个类型的 R 记为 $R-1$，其对应的数量记为 $N-1$；第2个类型的 R 记为 $R-2$，其对应的数量为 $N-2$，依此类推，则输出给云计算调度系统的预测结果为：

对应 $R-1$ 的 $N-1$，对应 $R-2$ 的 $N-2$，……，对应 $R-m$ 的 $N-m$。

云计算调度系统从中选择一种候选类型云资源进行调度这一过程，可具体细分为以下步骤：

步骤 S5-1，获取每一候选类型云资源的剩余资源量 STk 和单位成本 CTk；

步骤 S5-2，判断类型 T 云任务是否可以拆分并调度到不同云资源，如果不可以，则执行步骤 S5-3，如可以则执行步骤 S5-4；

步骤 S5-3，从所有候选类型云资源中选择一个剩余资源量大于或等于预测占用数量（预测占用数量等于步骤 S4 预测出的 T 类型云任务占用候选类型云资源的数量）且总成本最低的一个类型云资源进行调度。

例如：

对应 $R-1$ 的 $N-1$，$R-1$ 的剩余资源量为 $STk-1$，单位成本为 $CTk-1$；

对应 $R-2$ 的 $N-2$，$R-2$ 的剩余资源量为 $STk-2$，单位成本为 $CTk-2$；

……

对应 $R-m$ 的 $N-m$，$R-1$ 的剩余资源量为 $STk-m$，单位成本为 $CTk-m$。

如果 $N-1 \leqslant STk-1$，则计算选择 $R-1$ 的总成本 $(N-1)\times(CTk-1)$，否则不考虑对 $R-1$ 的选择；

如果 $N-2 \leqslant STk-2$，则计算选择 $R-2$ 的总成本 $(N-2)\times(CTk-2)$，否则不考虑对 $R-2$ 的选择；

……

如果 $N-m \leqslant STk-m$，则计算选择 $R-m$ 的总成本 $(N-m)\times(CTk-m)$，否则不考虑对 $R-m$ 的选择；

从所有计算的总成本中选出总成本最低的一个类型的云资源进行调度。

步骤 S5-4，令剩余数量 M' 的初始值为 M，从所有尚未被调度的候选类型云资源中选择有剩余资源量且单位成本最低的候选类型云资源，如果选择的候选类型云资源的剩余预测占用量（剩余预测占用数量＝预测占用数量 $\times(M'/M)$）小于或等于相应的剩余资源量，则将数量 M' 的 T 类型云任务全部调度到该候选类型云资源，如果选择的候选类型云资源的剩余预测占用数量大于相应的剩余资源量，则从数量 M' 的 T 类型云任务中拆分出 $[M'\times$（剩余资源量/剩余预测占用数量）]数量的 T 类型云任务调度到该候选类型云资源，并将 M' 更新为 $M'\times$（1－剩余资源量/剩余预测占用数量），重复这个步骤直至剩余资源量大于剩余预测占用数量。

例如：

对应 $R-1$ 的 $N-1$，$R-1$ 的剩余资源量为 $STk-1$，单位成本为 $CTk-1$；

对应 $R-2$ 的 $N-2$，$R-2$ 的剩余资源量为 $STk-2$，单位成本为 $CTk-2$；

……

对应 $R-m$ 的 $N-m$，$R-1$ 的剩余资源量为 $STk-m$，单位成本为 $CTk-m$。

按照单位成本从低到高排序。

令 $M'=M$

如果 $N-2 \leqslant STk-2$，则将数量 M' 的 T 类型云任务全部调度到 $R-2$ 类型云资源；

如果 $N-2 > STk-2$，则从数量 M' 的 T 类型云任务中拆分出 $[M'\times(STk-2/N-2)]$ 数量的 T 类型云任务调度到 $R-2$ 类型云资源，并将 M' 更新为 $M'\times(1-STk-2/N-2)$；接下来对 $CTk-4$ 进行类似处理，依此类推。

本具体技术方案通过步骤 S5-4 的这种拆分可以实现将数量 M 的 T 类型云任务调度

到多个候选类型云资源时的总成本最低。

(6) 步骤 S6,在数量的 T 类型云任务被调度且运行完成后,获取数量 M 的类型 T 云任务实际占用的被调度的候选类型云资源的数量,并存入过去能力大数据库。

本技术具有以下优点:

(1) 基于过去能力大数据和深度学习模型,通过过去不同类型的云任务调度到不同云资源上时对云资源的占用情况这一客观事实,来对调度任务中相应类型的云任务调度到相应类型云资源时将会占用的资源状况进行预测,与现有粗略估计的预估式预测方式相比,更加准确,使得基于本技术的预测结果的调度能最大程度地利用云资源,调度效率更高。

(2) 通过第一预设阈值、第二预设条件和第三预设条件确保了有足够的数据训练预设深度学习模型。

(3) 训练前先将检索出的每个过去调度记录按照时间先后顺序进行排序,令最近时间的训练数据对训练后的深度学习模型的影响更大,从而使得训练后的深度学习模型更能反映出最近的训练数据的特征。

(4) 在云计算调度完成后,本技术通过将实际占用的云资源的数量存入过去能力大数据库来修正和不断更新实际占用的云资源数量,使得预测结果更加准确和可靠。

5.1.2　基于大数据和深度学习模型的云计算调度系统

与图 5-1 的方法相对应,本技术基于大数据和深度学习模型的云计算调度系统,包括以下模块:

(1) 获取模块 1,用于获取需调度的云任务类型 T、需调度的 T 类型云任务的数量 M 以及拟调度到的云资源候选类型。

(2) 检索模块 2,用于从过去能力大数据库中检索出所有从类型 T 的云任务调度到候选类型的云资源上的调度记录,过去能力大数据库存储有过去每次调度中不同类型的云任务调度到不同类型的云资源上时云任务的类型、T 类型云任务的数量、云资源候选类型及候选类型云资源的占用数量。

(3) 预测模块 3,用于根据检索出的所有过去调度记录,采用深度学习模型预测调度数量 M 的 T 类型云任务到候选类型的云资源时会占用候选类型云资源的数量。

预测模块 3 包括:

初始化单元 3-1,用于初始化预设深度学习模型,使得预设深度学习模型的输入节点能以 T 类型云任务的数量作为输入,并使输出节点能以候选类型云资源的数量作为输出;

排序与训练单元 3-2,用于将检索出的每个过去调度记录按照时间先后顺序进行排序,并按照排序的先后依次作为训练数据对预设深度学习模型进行训练;

预测输出单元 3-3,用于将需调度的 T 类型云任务的数量 M 作为训练后的预设深度学习模型的输入,并将训练后的预设深度学习模型计算得到的输出作为预测出的调度数量 M 的 T 类型云任务到候选类型云资源时会占用候选类型云资源的数量。

(4) 调度模块 4,用于根据预测的结果完成云计算调度。

5.2 基于大数据和深度学习模型的停车引导

基于大数据和深度学习模型的停车引导的基本方法

传统的停车引导方法[21,22]，将符合用户要求的所有停车场中距离用户预设目的地最近的停车场推荐给用户，并且根据使用的地点不同，一般存在两种情况。

如果用户将车开到目的地后使用停车引导，如图 5-3 所示，那么从出发地到目的地的距离加上从目的地到被推荐的停车场的距离总会大于(大于的概率非常大)或等于从出发地到被推荐的停车场的距离(因为三角形的两边长之和肯定大于第三边长)，因此会增加用户的开车里程、延长用户的开车时间，从而造成用户成本的增加和用户时间的浪费。

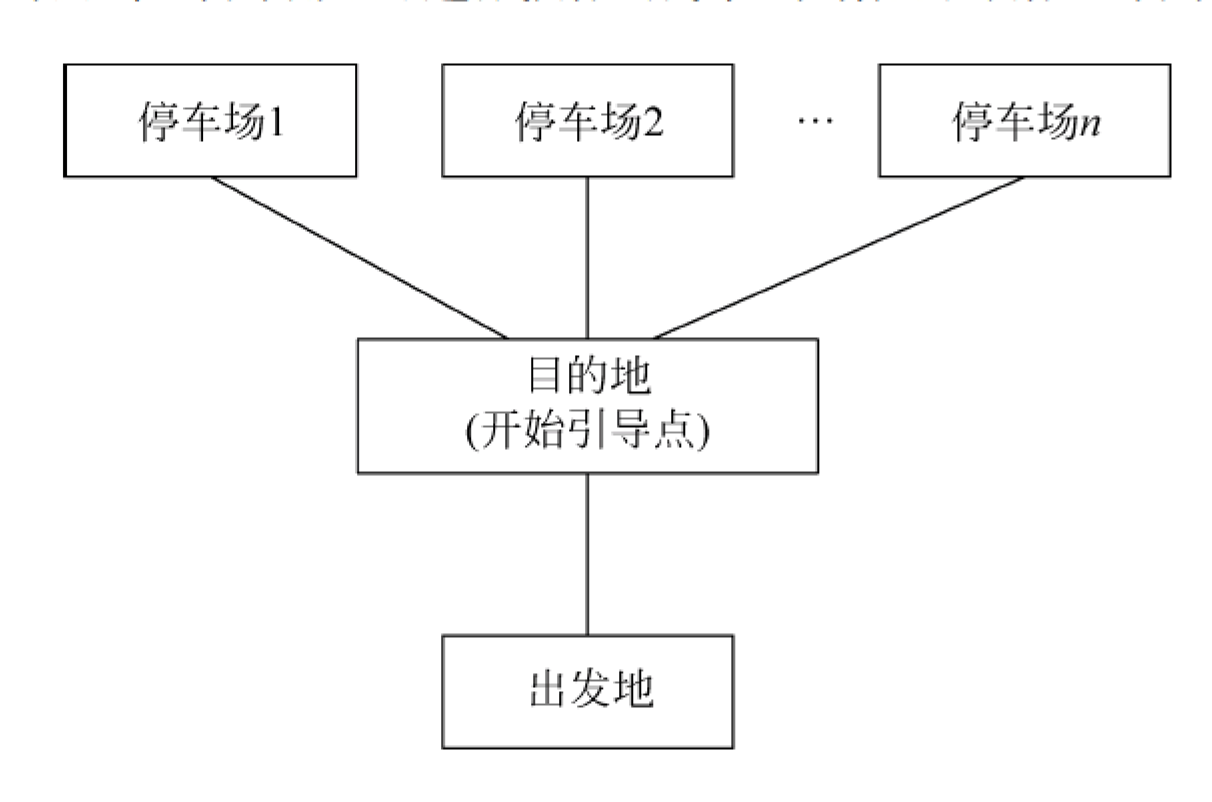


图 5-3　传统的停车引导方法示意图(一)

如果用户在出发地或距离目的地较远处使用停车引导，如图 5-4 所示，那么符合用户要求的停车场必然较远，因此会导致用户将车开到停车场需要较长时间，而在这较长时间内被推荐的停车场内的情况已经发生了变化(例如从有空闲停车位变为无空闲停车位)，那么就会导致用户停车失败。

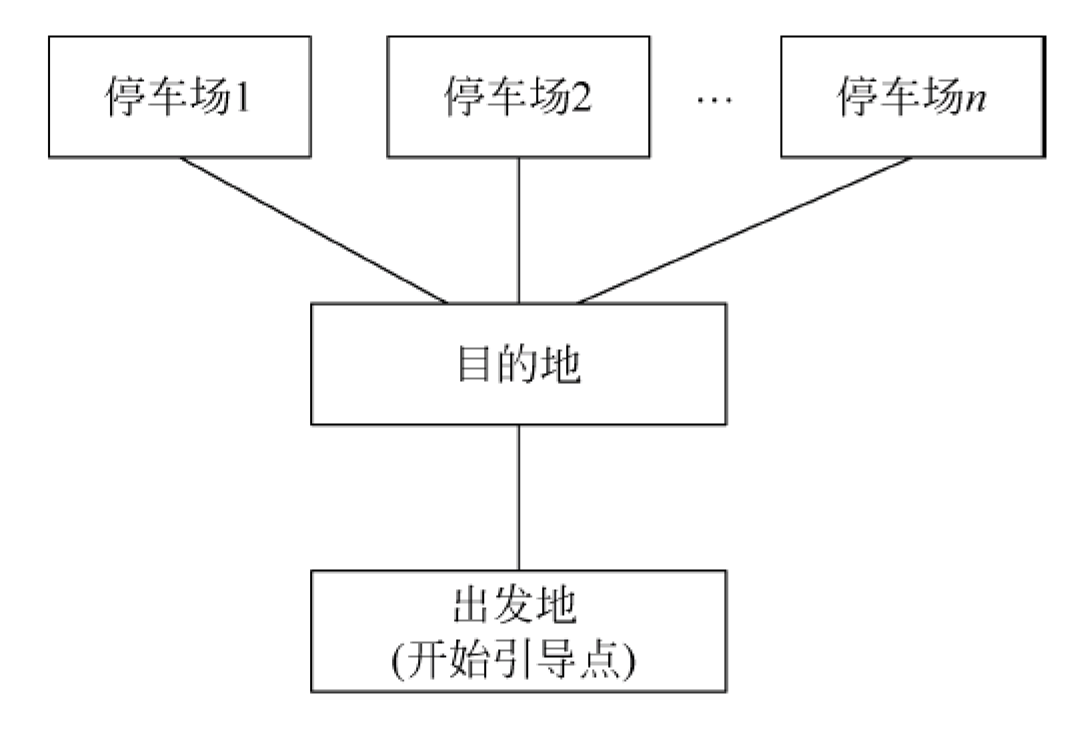


图 5-4　传统的停车引导方法示意图(二)

因此，传统的停车引导方法存在增加用户成本或者停车失败的问题。

本技术研究并给出了一种新的停车引导方法及系统，用于辅助用户选择目的地附近的停车场。方法包括：获取表示用户到达目的地时间的第一时间；获取预设的筛选条件；预

测在第一时间符合预设筛选条件的第一停车场集合；根据预测结果输出具有最优条件的停车场。系统包括时间获取模块、条件设置模块、预测模块、输出模块以及导航模块。上述方法及系统可以避免停车的人力和时间成本的浪费，又可避免停车失败的问题。

5.2.1　基于大数据和深度学习模型的停车引导方法

如图 5-5 所示为一技术方案的停车引导方法流程图，该方法用于辅助用户选择目的地附近的停车场，包括以下步骤：

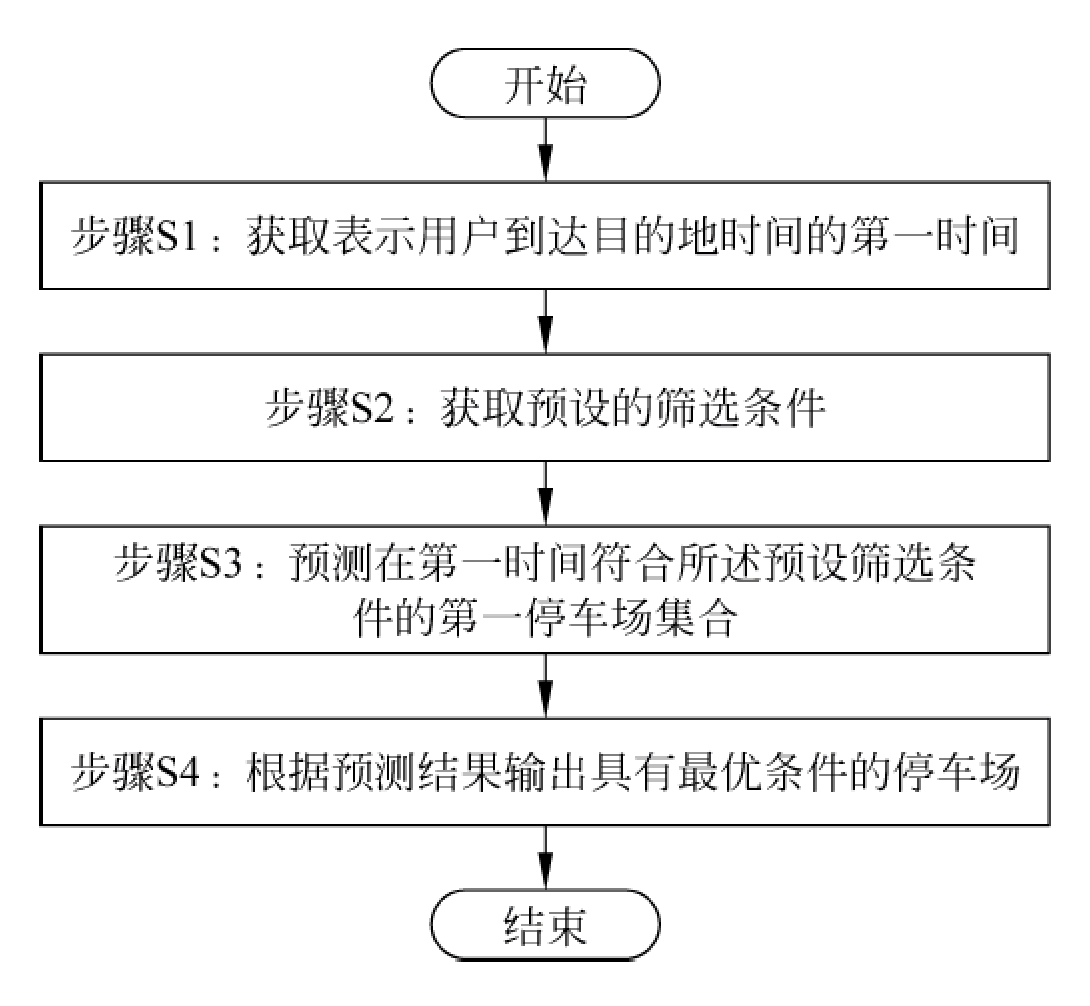


图 5-5　停车引导方法流程图

（1）步骤 S1，获取表示用户到达目的地时间的第一时间。第一时间不是用户开车实际到达的时间，而是预计可能到达目的地的时间，与实际到达的时间存在一定的差别，但一般会在合理的误差区间内。预估第一时间的目的，是预测在第一时间时停车场的车位状况。

（2）步骤 S2，获取预设的筛选条件。预设的筛选条件可以用来筛选目的地附近的停车场，一般是符合用户要求的各种条件。筛选条件可以包括距离约束条件和车位空闲状况约束条件。其中距离约束条件是指以目的地为中心、约束距离为半径的范围。车位空闲状况约束条件是指车位数量、比例等。

筛选条件可以由用户自行设置，例如用户选择目的地方圆 500m 内的停车场、停车位数量在 20 以上等。当用户没有设置筛选条件时，调取默认条件，例如目的地方圆 500m 内的停车场、停车位数量在 20 以上等。

（3）步骤 S3，预测在第一时间符合预设筛选条件的第一停车场集合。预测目的地附近的各停车场的车位变化，并根据预设的筛选条件将车位状况变化后的停车场进行筛选，获得第一停车场集合。

（4）步骤 S4，根据预测结果输出具有最优条件的停车场。从第一停车场集合中选择一个具有最优条件的停车场输出。该最优条件可以是距离近、车位多、环境好等条件中的一种或多种条件的综合评价。输出方式可以是向用户推荐，例如，输出第一停车场集合中距离目的地最近的停车场。

上述停车引导方法,通过预估车辆到达目的地的第一时间,并预测在第一时间时各停车场的车位变化,根据预测结果推荐停车场,进而可以利用导航系统规划从当前位置到达具有最优条件的停车场的路线。该方法既可以提前规划到达停车场的路线,避免人力和时间成本的浪费,又相对准确,在预测算法稳定时,可以大概率避免因提前设定停车场导致停车失败的问题。

基于大数据和深度学习模型的停车引导的详细方法

在步骤 S1 中,当用户比较熟悉目的地,知道从出发地到目的地一般花多长时间,则可以由用户提供该第一时间;当用户没有提供第一时间,则按照如下方法计算第一时间,如图 5-6 所示。

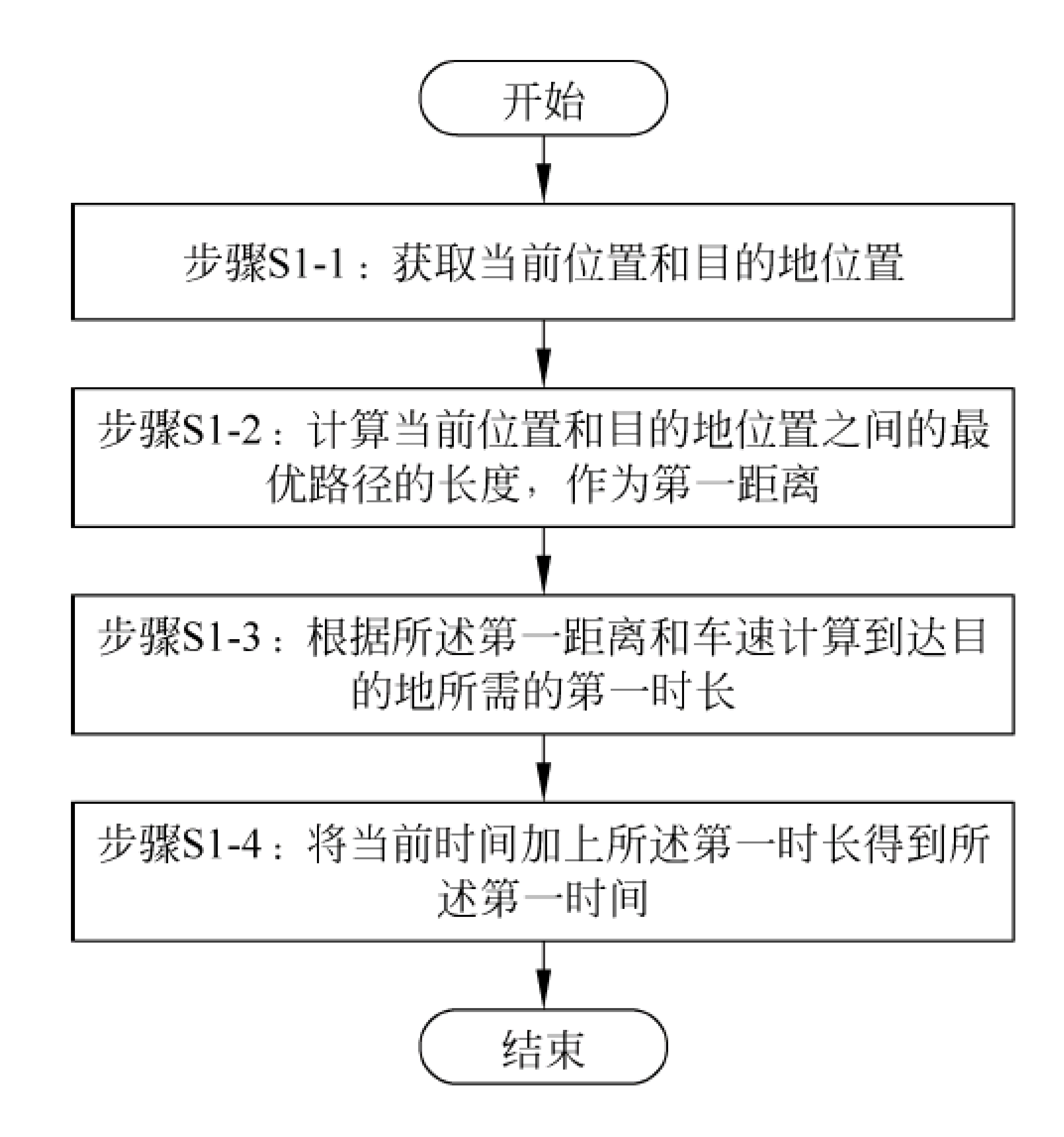


图 5-6　获取第一时间的方法流程图

步骤 S1-1,获取当前位置和目的地位置。获取当前位置的方法为用户输入或定位(例如 GPS 定位)获得,目的地位置为用户预设。

步骤 S1-2,计算当前位置和目的地位置之间最优路径的长度,作为第一距离。可以由车载导航系统获得该最优路径。

步骤 S1-3,根据第一距离和车速计算到达目的地所需的第一时长。车速可以是一段时间以内的平均车速,也可以是结合其他情况计算所得,例如根据路况等。

步骤 S1-4,将当前时间加上第一时长得到第一时间。当前时间即选择停车场的时间。

当开始采用上述技术方案的方法选择停车场时,就开始执行上述处理流程,以处理上述流程的开始时间为当前时间。

如图 5-7 所示,上述步骤 S3 具体可以采用如下处理过程:

步骤 S3-1,根据距离约束条件筛选出第二停车场集合。即搜索并筛选在距离范围内的所有停车场,作为第二停车场集合。

步骤 S3-2,将第二停车场集合中、当前时间每一停车场的状况输入对应停车场的已训练深度学习模型进行深度学习,得到每一停车场的已训练深度神经网络的输出。即通过已训练学习模型网络,根据当前停车场的状况预测第一时长后(即第一时间)的停车场的状况。

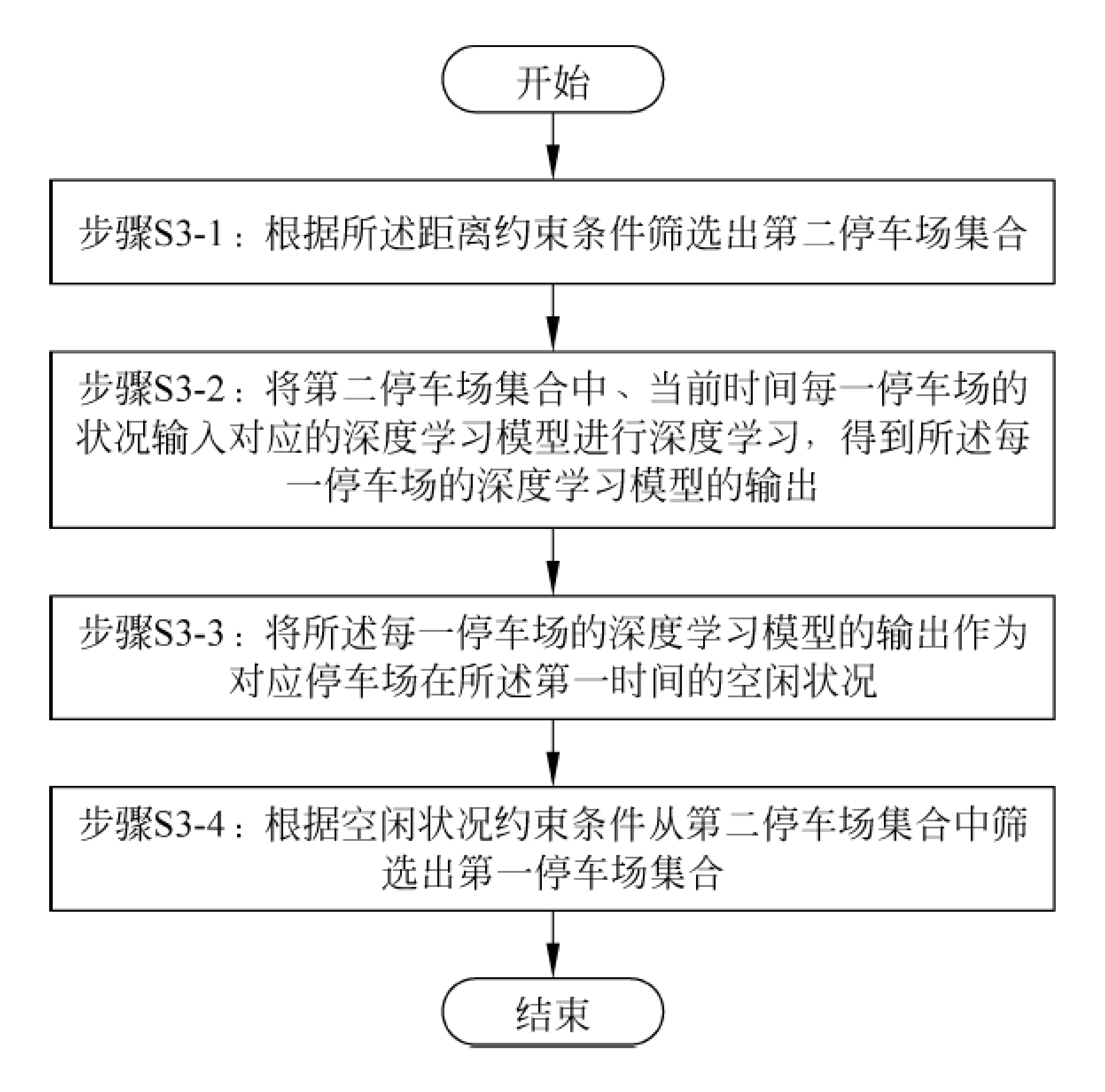


图 5-7　预测过程的具体方法流程图

步骤 S3-3，将每一停车场的已训练深度学习模型的输出作为对应停车场在第一时间的空闲状况。空闲状况可以采用空闲车位的比例、数量或根据比例、数量定义的预设状态来表示。例如，0%表示完全空闲，50%表示一半空闲，100%表示零空闲；停车场的空闲状况也可以用"非常空闲""比较空闲""不空闲"等预设状态来表示，为了便于计算，这些预设状态可以对应为数字编码，例如"非常空闲"对应 1、"比较空闲"对应 2、"不空闲"对应 3 等。

步骤 S3-4，根据空闲状况约束条件从第二停车场集合中筛选出第一停车场集合。

从第二停车场集合中筛选出第一时间的空闲状况符合空闲状况约束条件的停车场加入第一停车场集合。

由于深度学习模型的预测精度非常高(这已经在很多应用中得到了印证，例如图像识别、语音识别，而空闲状况与图像数据或语音数据没有本质上的区别。例如将停车场中每个车位作为一个像素点，如果占用则该像素点为黑，否则为白，则每个时间的停车场的空闲状况都可以用一个图像来表示)，所以可以高精度地预测到停车场在车子到达预设目的地时的空闲状况。

在预测步骤 S3-2 之前，包括步骤 S3-2′：采用第二停车场集合中、具有第一时长跨度的每一停车场空闲状况数据训练对应停车场的深度学习模型，得到每一停车场的深度学习模型。深度学习模型一般需要输入大量有效的数据对其进行训练，才能在预测时更加准确。本技术方案中，由于需要预测第一时长后的车位状况变化，因此需要输入具有第一时长跨度的停车场空闲状况数据对停车场的深度神经网络进行训练。

如图 5-8 所示，采用具有第一时长跨度的每一停车场的空闲状况数据训练一个深度学习模型，得到每一停车场的已训练深度神经网络，具体可以包括以下步骤：

步骤 S3-2-1′，初始化一个具有预设层数个中间层的深度学习模型。

将深度学习模型的输入层及输出层的数据结构初始化为每一停车场的空闲状况数据的

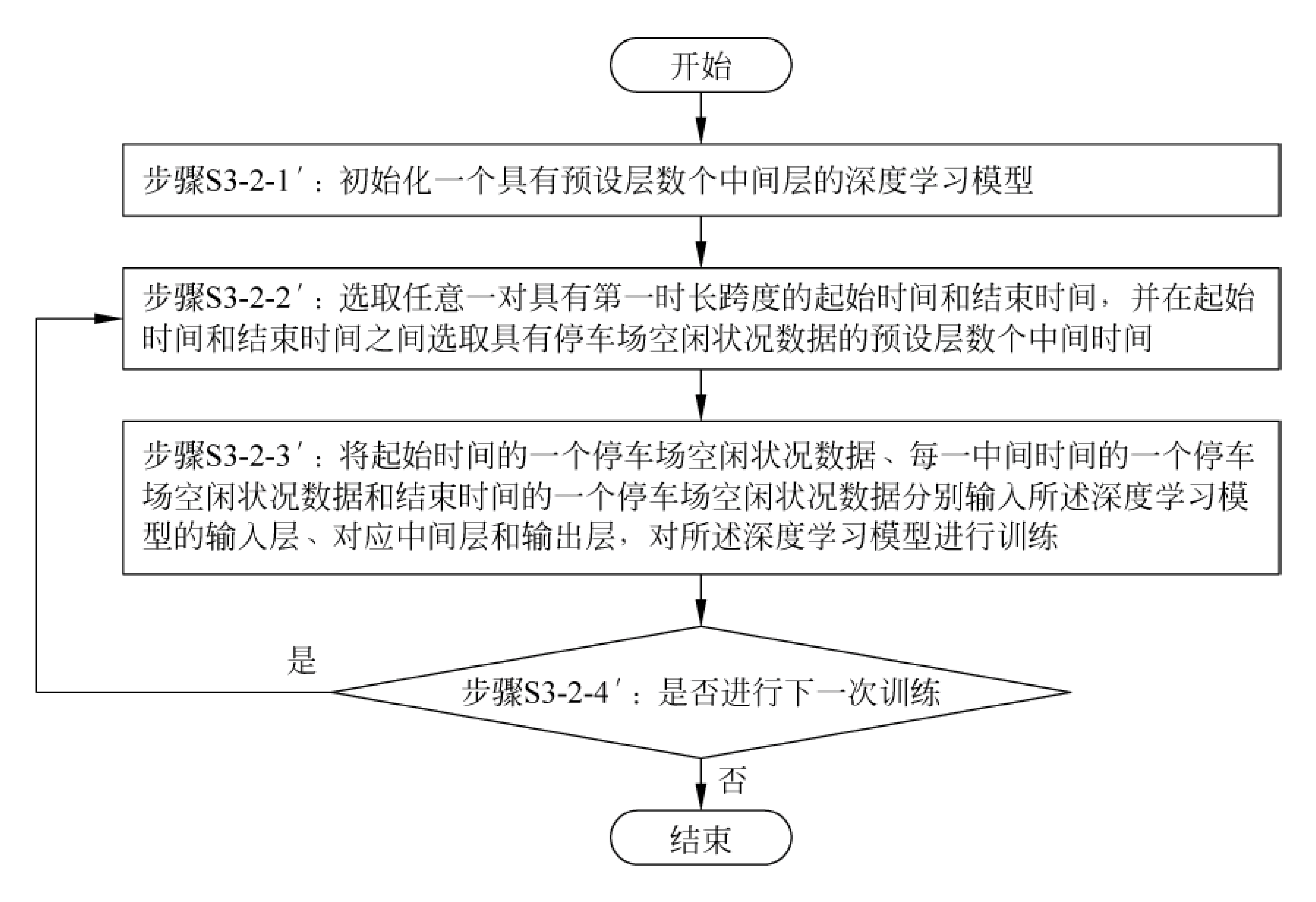


图 5-8　训练深度学习模型的方法流程图

数据结构。或者,也可以将深度学习模型的输入层及输出层的数据结构初始化为每一停车场的空闲状况数据可以转化成的数据结构。同时,将深度学习模型初始化出预设层数个中间层。中间层的预设层数可以是任一大于或者等于 0 且小于或者等于第一时长跨度中除起始时间和结束时间之外的具有停车场空闲状况数据的时间的个数。

步骤 S3-2-2′,选取任意一对具有第一时长跨度的起始时间和结束时间,并在起始时间和结束时间之间选取具有停车场空闲状况数据的预设层数个中间时间。选取的一对起始时间和结束时间可以是历史上任意的时间,并且在该时间有相关的停车场空闲状况数据。优选地,可以选取历史上每一天与本次预测对应的出发时间和到达时间。

步骤 S3-2-3′,将起始时间的每一停车场的空闲状况数据、每一中间时间的每一停车场的空闲状况数据和结束时间的每一停车场的空闲状况数据分别输入深度学习模型的输入层、对应中间层和输出层,对深度学习模型进行训练。将每一停车场历史上的实际数据输入到深度学习模型中进行训练,可以获得每一停车场的深度学习模型。训练的数据越有效且训练数据量越大,预测的结果越准确。

步骤 S3-2-4′,判断是否对深度学习模型进行下一次训练,若是,则回到步骤 S3-2-2′,重新执行上述各步骤,否则结束并将深度学习模型作为每一停车场的已训练深度学习模型。在执行此次预测之前,都可以利用每一停车场的已有数据持续对深度学习模型进行训练,直到利用验证集对深度学习模型进行验证时得到的误差小于阈值为止。

判断的步骤是:

选取距离当前时间最近的一对具有第一时长跨度的起始时间和结束时间;

将起始时间的每一停车场的空闲状况数据输入深度学习模型的输入层,然后通过深度学习模型的深度学习得到深度学习模型的输出层的输出结果;

将输出结果与结束时间的每一停车场的空闲状况数据进行对比,如果对比得到的误差大于预设阈值,则对深度学习模型进行下一次训练,否则结束训练并将深度学习模型作为每

一停车场的已训练深度学习模型。

在预测步骤 S3-2 之前，深度学习模型也可以是已经训练成型，可以直接在步骤 S3-2 中使用。

上述训练过程中所涉及的数据，即起始时间的停车场空闲状况数据和结束时间的停车场空闲状况数据都存入大数据，并且还将各停车场的实时空闲状况数据都存入大数据。这些大量的数据可以用作大数据分析，进一步帮助提高预测的准确性。

5.2.2　基于大数据和深度学习模型的停车引导系统

1. 基于大数据和深度学习模型的停车引导系统的基本模块

如图 5-9 所示，停车引导系统包括时间获取模块 1、条件设置模块 2、预测模块 3、输出模块 4 以及导航模块 7。

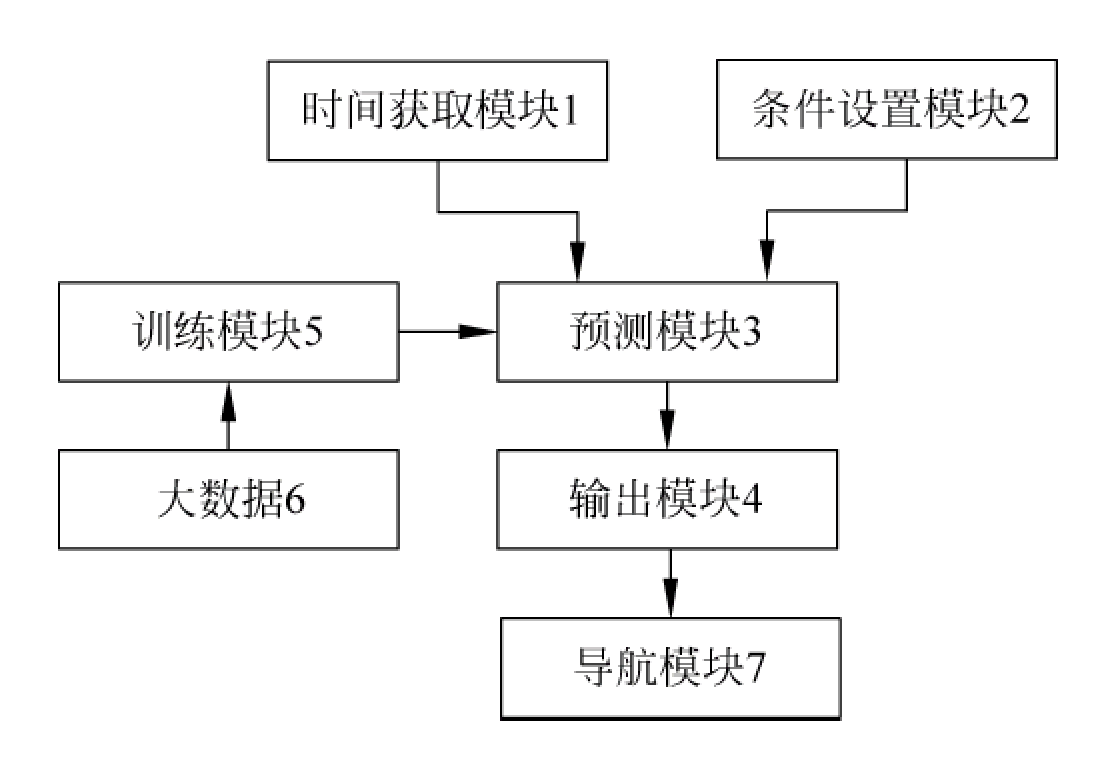


图 5-9　停车引导系统模块图

(1) 时间获取模块 1，用于获取表示用户到达目的地时间的第一时间。第一时间不是用户开车实际到达的时间，而是预计可能到达目的地的时间，与实际到达的时间存在一定的差别，但一般会在合理的误差区间内。预估第一时间的目的，是预测在第一时间时停车场的车位状况。

(2) 条件设置模块 2，用于获取预设的筛选条件。预设的筛选条件可以用来筛选目的地附近的停车场，筛选条件一般是符合用户要求的各种条件。筛选条件可以包括距离约束条件和车位空闲状况约束条件。其中距离约束条件是指以目的地为中心、约束距离为半径的范围。车位空闲状况约束条件是指车位数量、比例等满足一定的条件。

筛选条件可以由用户自行设置，例如用户选择目的地方圆 500m 内的停车场、停车位数量在 20 以上等。当用户没有设置筛选条件时，调取默认条件。

(3) 预测模块 3，用于预测在第一时间符合预设筛选条件的第一停车场集合。

(4) 输出模块 4，根据预测结果输出具有最优条件的停车场。从第一停车场集合中选择一个具有最优条件的停车场输出。该最优条件可以是距离近、车位多、环境好等条件中的一种或多种条件的综合评价。

(5) 导航模块 7，利用导航系统规划从当前位置到达具有最优条件的停车场的路线。

上述停车引导系统，通过预估车辆到达目的地的第一时间，并预测在第一时间时各停车

场的车位变化，根据预测结果推荐停车场，进而可以利用导航系统规划从当前位置到达具有最优条件的停车场的路线。该方法既可以提前规划到达停车场的路线，避免人力和时间成本的浪费，又相对准确，在预测算法稳定时，可以大概率避免因提前设定停车场导致停车失败的问题。

2. 基于大数据和深度学习模型的停车引导系统的扩展模块

时间获取模块 1 可以包括供用户输入第一时间的输入单元，由用户提供该第一时间。例如，用户比较熟悉目的地，知道从出发地到目的地一般花多长时间，则可以由用户提供该第一时间。

如图 5-10 所示，时间获取模块 1 也可以包括位置获取单元 1-1、距离计算单元 1-2、时长计算单元 1-3 以及时间计算单元 1-4。

位置获取单元 1-1，用于获取当前位置和目的地位置。获取当前位置的方法为用户输入或定位获得，目的地位置为用户预设。

距离计算单元 1-2，用于计算当前位置和目的地位置之间最优路径的长度，作为第一距离。

时长计算单元 1-3，用于根据第一距离和车速计算到达目的所需的第一时长。

时间计算单元 1-4，用于将当前时间加上第一时长得到第一时间。

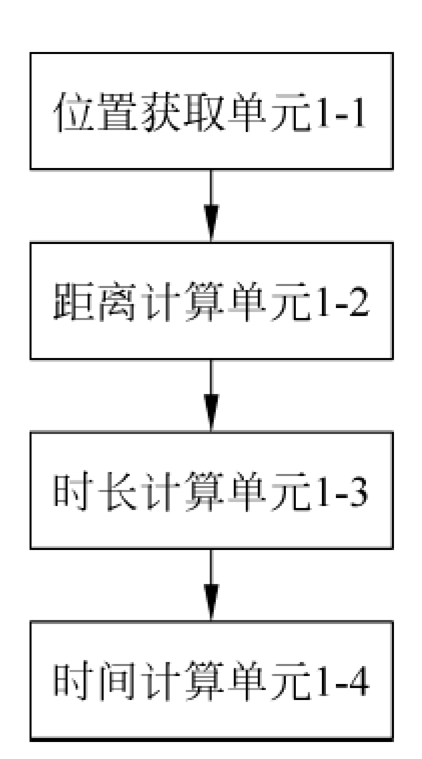


图 5-10　时间获取模块的一种具体模块图

如图 5-11 所示，条件设置模块 2 可以包括用户设置单元 2-1 和系统设置单元 2-2。用户设置单元 2-1 供用户设置预设的筛选条件。系统设置单元 2-2 设置系统默认条件，用于当用户没有设置筛选条件时供调取。

如图 5-12 所示，预测模块 3 包括依距离筛选单元 3-1、深度神经网络处理单元 3-2 以及依空闲状况筛选单元 3-3。

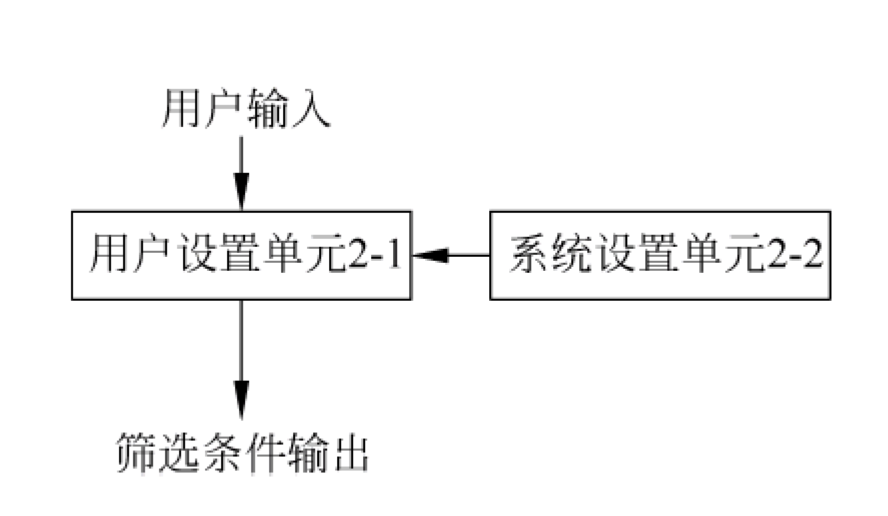


图 5-11　条件设置模块的一种具体模块图

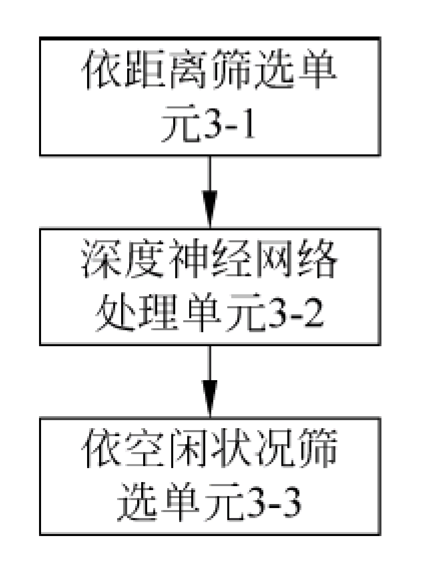


图 5-12　预测模块的一种具体模块图

依距离筛选单元 3-1，用于根据距离约束条件筛选出第二停车场集合。

深度神经网络处理单元 3-2，用于将第二停车场集合中、当前时间每一停车场的状况输入对应停车场的已训练深度学习模型进行深度学习，得到每一停车场的已训练深度学习模型的输出；并将每一停车场的已训练深度学习模型的输出作为对应停车场在第一时间的空闲状况。空闲状况采用空闲车位的比例、数量或根据比例、数量定义的预设状态来表示。

依空闲状况筛选单元 3-3，用于根据空闲状况约束条件从第二停车场集合中筛选出第一停车场集合。即从第二停车场集合中筛选出第一时间的空闲状况符合空闲状况约束条件的停车场加入第一停车场集合。

停车引导系统还可以包括训练模块 5。训练模块 5 用于采用第二停车场集合中、具有第一时长跨度的每一停车场空闲状况数据训练对应停车场的深度学习模型，得到每一停车场的深度学习模型。如图 5-13 所示，训练模块 5 包括选择单元 5-1、训练单元 5-2 以及循环控制单元 5-3。

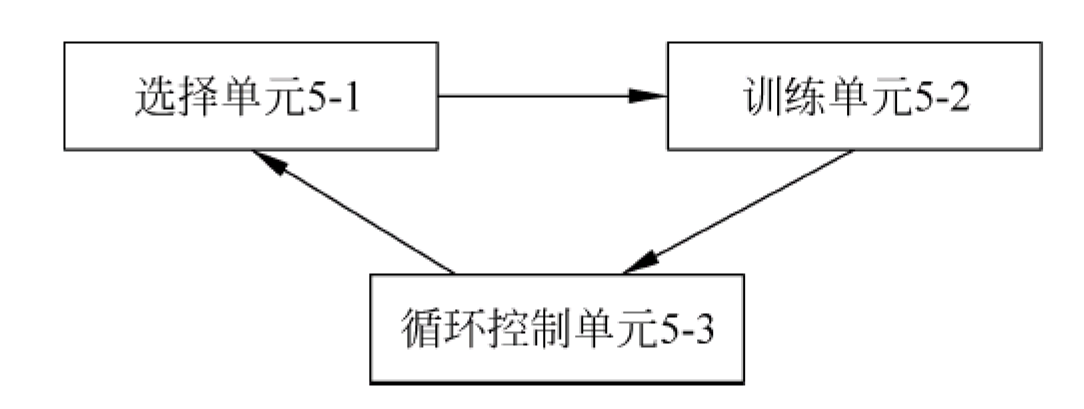


图 5-13 训练模块的一种具体模块图

选择单元 5-1，用于选取任意一对具有第一时长跨度的起始时间和结束时间。

训练单元 5-2，将起始时间的每一停车场空闲状况数据和结束时间的每一停车场空闲状况数据输入每一停车场的深度学习模型进行训练。

循环控制单元 5-3，用于控制选择单元 5-1 和训练单元 5-2 重复工作。

停车引导系统还可以包括大数据 6，用于存储起始时间的停车场空闲状况数据和结束时间的停车场空闲状况数据。上述训练过程中所涉及的数据，即起始时间的停车场空闲状况数据和结束时间的停车场空闲状况数据都存入大数据，并且还将各停车场的实时空闲状况数据都存入大数据。这些大量的数据可以用作大数据分析，进一步帮助提高预测的准确性。

5.3 基于大数据与深度学习的婚恋对象匹配推荐

大数据指需要新处理模式才能具有更强的决策力、洞察发现力以及流程优化能力的海量、高增长率和多样化的信息资产[24,25]。在日常生活中发现，夫妻往往是有夫妻相的，因为人的长相是与基因相关的，而基因与人的性格等各方面相关。

现有婚介机构或网站向用户随机或根据红娘的经验向用户推荐婚恋对象，没有客观依据，使得推荐的婚恋对象与用户的匹配度低。

5.3.1 基于大数据与深度学习的婚恋对象匹配推荐方法

1. 基于大数据与深度学习的婚恋对象匹配推荐的基本方法

如图 5-14 所示，在一个技术方案中，提出了一种婚恋对象匹配数据处理方法，包括以下步骤：

（1）步骤 S1，获取待匹配的第一和第二对象图像数据。

其中，第一、第二对象图像数据是分别表征第一和第二对象特征的数据，可以是拍摄的图片，也可以是对图片进行图像特征提取后的图像特征数据，其中第一和第二对象分别为请求匹配对象和待推荐对象。第一、第二对象图像数据可以携带在终端发送的匹配请求中，也可以是从服务器获取的第一、第二对象图像数据。在一个技术方案中，第二对象图像数据是从服务器存储的注册用户对应的图像信息提取的图像数据，随着注册用户信息更新。在一个技术方案中，第一和第二对象图像数据是分别从第一和第二对象图像中提取人脸区域得到的。

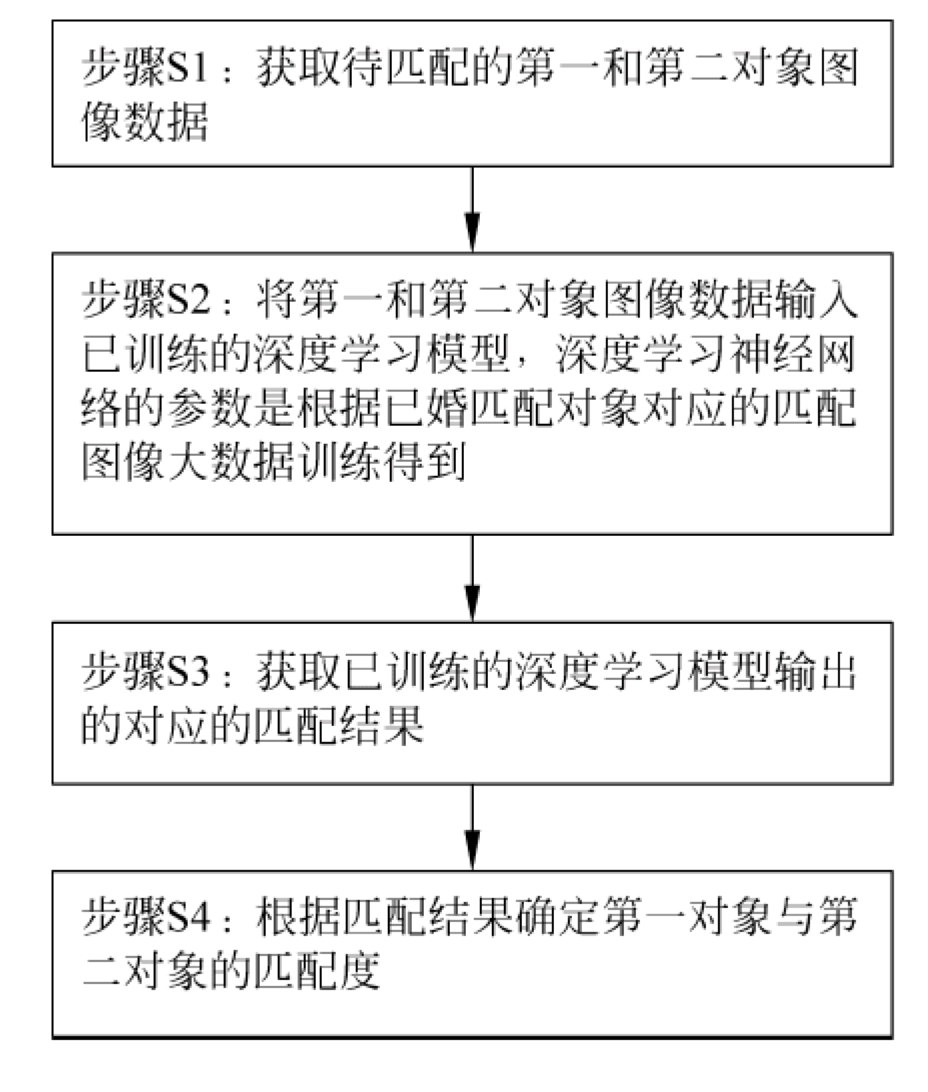


图 5-14　婚恋对象匹配数据处理方法的流程图

（2）步骤 S2，将第一和第二对象图像数据输入已训练的深度学习模型，深度学习模型的参数是根据已婚匹配对象对应的匹配图像大数据训练得到。

深度学习模型是模拟人脑进行分析学习的神经网络，模仿人脑的机制来解释图像数据，深度学习的第一步实质上是一个网络参数初始化过程，区别于传统神经网络初值随机初始化，深度学习模型是通过无监督训练输入数据的结构得到的，因而这个初值更接近全局最优，从而能够取得更好的效果。在一个技术方案中，将卷积神经网络作为此深度学习模型，卷积神经网络（Constitutional Neural Networks，CNN）是在多层神经网络的基础上发展起来的针对图像分类和识别而特别设计的一种深度学习方法。

已训练的深度学习模型是与待匹配的对象的数据格式和输出格式匹配度最高的。已训练的深度学习模型包括输入层、多个隐藏层和输出层，其中层数、每层预设节点数、各网络连接的权值由训练得到。在一个技术方案中，根据已训练的深度学习模型的预设数据格式对请求匹配对象图片及每一候选待推荐对象图片进行预处理，具体包括：判断请求匹配对象图片及每一候选待推荐对象图片中的各个图片与预设数据格式是否一致，否则将图片的格式转化为预设数据格式。

已婚匹配对象是指法定效力内的结婚登记对象，获取数据量足够大的已婚匹配对象对应的匹配图像大数据作为深度学习模型的训练数据进行训练，得到深度学习模型的参数。在一个技术方案中，匹配对象对应的匹配图像大数据包括已婚匹配对象双方的人脸图片，还可以包括双方的地域、结婚登记地域、年龄信息形成多维度的输入训练数据进行训练。可针对不同结婚登记地区生成对应的不同的已训练的深度学习模型，提高不同地区对象匹配的适应度与准确性。由于法定效力内的结婚登记对象代表匹配成功，根据匹配成功对象的图像信息，形成客观的科学依据大数据进行深度网络分析，提高了匹配对象的客观依据性。

（3）步骤 S3，获取已训练的深度学习模型输出的对应的匹配结果。

深度学习模型输出的匹配结果形式不限，可以是匹配概率，也可以是互斥的匹配标签，如用“1”表示匹配，用“0”表示不匹配。已训练的深度学习模型可以根据输入的待匹配对象的图像数据分析输出对应的客观的匹配结果。

（4）步骤 S4，根据匹配结果确定第一对象与第二对象的匹配度。

如果匹配结果是匹配概率，则可根据匹配概率的大小确定第一对象与第二对象的匹配度。如果匹配结果是互斥的匹配标签，则直接根据匹配标签的值得到对应的匹配度，如匹配成功或匹配失败。可将第一对象与多个待匹配对象进行匹配度的计算，获取最大匹配概率的对象作为目标匹配对象。

本技术方案中，通过获取待匹配的第一和第二对象图像数据，将第一和第二对象图像数据输入已训练的深度学习模型，深度学习模型的参数是根据已婚匹配对象对应的匹配图像大数据训练得到，获取已训练的深度学习模型输出的对应的匹配结果，根据匹配结果确定第一对象与第二对象的匹配度，基于大数据和深度学习模型建立待匹配对象与匹配度的相关性，提供客观的匹配依据。

2. 基于大数据与深度学习的婚恋对象匹配推荐的详细方法

在一个技术方案中，如图 5-15 所示，步骤 S1 之前，包括以下步骤：

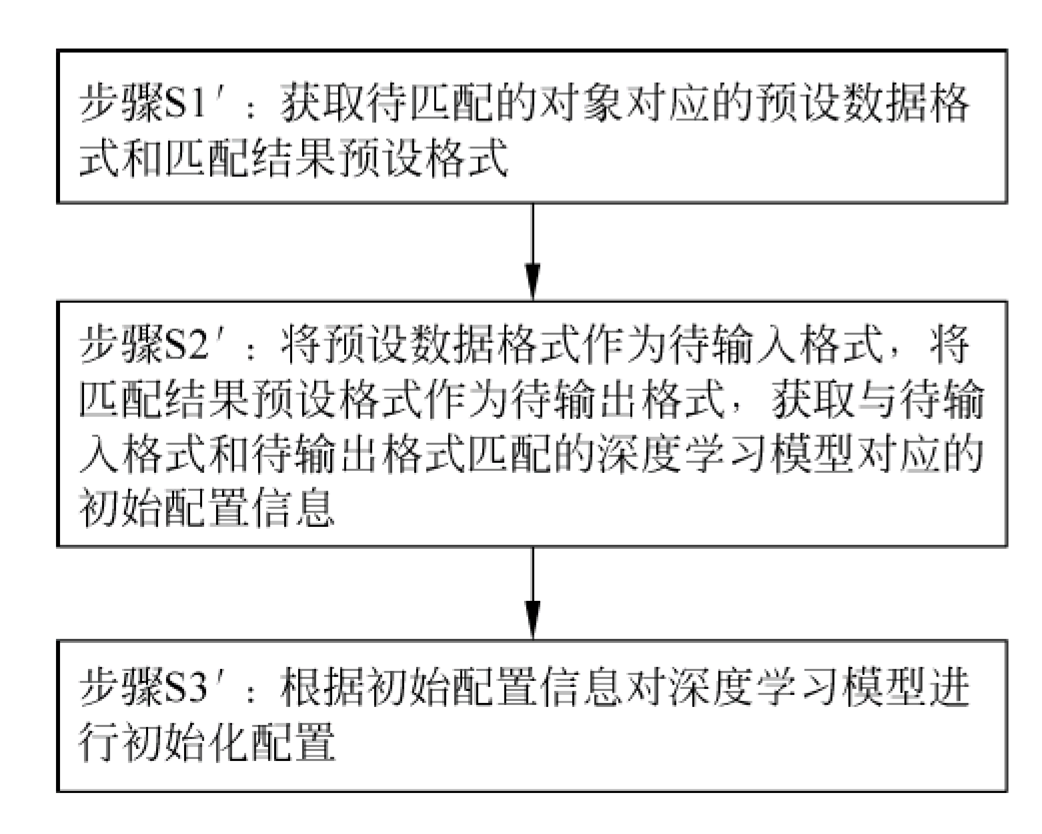


图 5-15 初始化配置的流程图

步骤 S1′，获取待匹配的对象对应的预设数据格式和匹配结果预设格式。

预设数据格式可以通过预先设定，也可以通过如下步骤自动获取：从待匹配的对象对应的每一对请求匹配对象照片和候选待推荐对象照片中每一照片的格式信息中获取数据格式，如 1080 像素×1920 像素，统计获取到的不同数据格式中每一数据格式的照片的数量，获取最大的数量对应的数据格式，作为预设数据格式。匹配结果预设格式是指需要得到的匹配结果的格式，如果是需要得到概率，则预设格式为 0～1 的小数，如果是需要得到标签，则预设格式为标签对应的格式。

步骤 S2′，将预设数据格式作为待输入格式，将匹配结果预设格式作为待输出格式，获取与待输入格式和待输出格式匹配的深度学习模型对应的初始配置信息。

深度学习模型知识库中预先存有各类输入格式和输出格式对应的深度学习模型配置信息，需要从中筛选出与当前待输入格式和待输出格式匹配的深度学习模型对应的初始配置信息。具体的筛选方式可根据需要自定义，如计算当前待输入格式与深度学习模型的候选输入格式的输入匹配度，计算当前待输出格式与候选输出格式的输出匹配度，然后根据输入匹配度与输出匹配度综合计算最终的匹配度。将匹配度最大的候选输入格式与候选输出格式对应的深度学习模型配置信息作为最终的初始配置信息。

配置信息是指用于配置深度学习模型的信息，包括预设层数、每层预设节点数、各网络连接的预设权值等。

步骤 S3′，根据初始配置信息对深度学习模型进行初始化配置。

将初始配置信息中的各个配置项分别赋予深度学习模型中对应的项，从而得到初始化的深度学习模型。可以接受修改请求，修改请求携带修改配置项和对应的修改值，从而根据修改值对初始化的深度学习模型的修改配置项进行修改。

本技术方案中，通过预设数据格式和匹配结果预设格式筛选得到匹配度高的深度学习模型对应的初始配置信息，能提高初始配置信息的准确度。

在一个技术方案中，如图 5-16 所示，步骤 S1 之前，包括以下步骤：

步骤 S1″，获取训练图像大数据，训练图像大数据包括已婚匹配对象对应的匹配图像大数据。

训练图像大数据是指用于训练深度学习模型的海量数据，其中已婚匹配对象对应的匹配图像大数据包括多个不同地区的已婚匹配对象双方对应的图像数据，如已婚对象双方的图片。可以针对不同地区分别采用对应地区的训练图像大数据训练得到与地区对应的深度学习模型，因为不同的地区由于地域、人口、风俗文化的差异在对象匹配度上也存在差异，针对不同地区形成不同的深度学习模型有利于提高当地的匹配客观性与准确率。

步骤 S1″：获取训练图像大数据，训练图像大数据包括已婚匹配对象对应的匹配图像大数据

步骤S2″：从匹配图像大数据中提取对应的各个已婚匹配对象对应的匹配图像对，组成匹配图像对集合

步骤S3″：将任意两个已婚匹配图像对进行拆分组合形成非匹配图像对，各个非匹配图像对组成非匹配图像对集合

步骤S4″：将各个匹配图像对和各个非匹配图像对分别作为深度学习模型的输入，对深度学习模型进行无监督训练

步骤S5″：将各个匹配图像对作为深度学习模型的输入，将匹配图像对对应的预设匹配标签作为深度学习模型的预期输出，对深度学习模型进行有监督训练

步骤S6″：将各个非匹配图像对作为深度学习模型的输入，将非匹配图像对对应的预设非匹配标签作为深度学习模型的预期输出，对深度学习模型进行有监督训练

步骤S7″：得到已训练的深度学习模型

图 5-16 训练深度学习模型的流程图

步骤 S2″，从匹配图像大数据中提取对应的各个已婚匹配对象对应的匹配图像对，各个已婚匹配对象对应的匹配图像对组成匹配图像对集合。

匹配图像对中是指已婚匹配对象双方的图像，可为人脸图像、全身图像等，根据需要自定义需要的图像范围。可通过人脸识别算法从中提取人脸得到各个已婚匹配对象对应的匹配图像对。如已婚匹配对象包括 A、B、C、D，则匹配图像对分别为$\{A_1, A_2\}$，$\{B_1, B_2\}$，$\{C_1, C_2\}$，$\{D_1, D_2\}$，匹配图像对集合 $P=\{(A_1, A_2), (B_1, B_2), (C_1, C_2), (D_1, D_2)\}$。

在一个技术方案中，匹配图像大数据是结婚证数据，从结婚证中根据性别特征和人脸特征进行图像识别，提取丈夫照片和妻子照片形成各个已婚匹配对象对应的匹配图像对；也可以利用丈夫照片和妻子照片在结婚证中的预设位置，来根据位置范围对丈夫照片和妻子照片进行直接提取，因为结婚证都是标准的，丈夫照片在结婚证的固定位置范围内，妻子照片也在结婚证的另一固定位置范围内，可快速提取得到匹配图像对，提高数据准备的效率。

步骤 S3″，将任意两个已婚匹配图像对进行拆分组合形成非匹配图像对，各个非匹配图像对组成非匹配图像对集合。

将一个匹配图像对中的图像与任意另一个匹配图像对中相同属性的图像进行交换，相同属性是指相同性别，从而形成非匹配图像对，各个非匹配图像对组成非匹配图像对集合。如匹配图像对集合 $P=\{(A_1,A_2),(B_1,B_2),(C_1,C_2),(D_1,D_2)\}$，形成非匹配图像对集合 $Q=\{(A_1,B_2),(A_1,C_2),(A_1,D_2),(B_1,A_2),(B_1,C_2),(B_1,D_2),(C_1,A_2),(C_1,B_2),(C_1,D_2),(D_1,A_2),(D_1,B_2),(D_1,C_2)\}$。

步骤 S4″，将各个匹配图像对和各个非匹配图像对分别作为深度学习模型的输入，对深度学习模型进行无监督训练。

将各个匹配图像对，如 (A_1,A_2)，(B_1,B_2)，(C_1,C_2)，(D_1,D_2) 分别输入深度学习模型，进行无监督训练，将各个非匹配图像对，如 (A_1,B_2)，(A_1,C_2)，(A_1,D_2)，(B_1,A_2)，(B_1,C_2)，(B_1,D_2)，(C_1,A_2)，(C_1,B_2)，(C_1,D_2)，(D_1,A_2)，(D_1,B_2)，(D_1,C_2) 分别输入深度学习模型，进行无监督训练。采用自下而上的无监督训练，逐层构建单层神经元，每层采用 Wake-Sleep 算法进行调优。每次仅调整一层，逐层调整，这个过程可以看作是一个 Feature Learning 的过程，是和传统神经网络区别较大的部分。Wake-Sleep 算法分为 Wake 阶段和 Sleep 阶段，其中 Wake 阶段是认知过程，通过下层的输入特征（Input）和向上的认知（Encoder）权重产生每一层的抽象表示（Code），再通过当前的生成（Decoder）权重产生一个重建信息（Reconstruction），计算输入特征和重建信息残差，使用梯度下降修改层间的下行生成（Decoder）权重，也就是"如果现实跟我想象的不一样，改变我的生成权重使得我想象的东西变得与现实一样"。Sleep 阶段是生成过程，通过上层概念（Code）和向下的生成（Decoder）权重，生成下层的状态，再利用认知（Encoder）权重产生一个抽象景象。利用初始上层概念和新建抽象景象的残差，利用梯度下降修改层间向上的认知（Encoder）权重，也就是"如果梦中的景象不是我脑中的相应概念，改变我的认知权重使得这种景象在我看来就是这个概念"。

步骤 S5″，将各个匹配图像对作为深度学习模型的输入，将匹配图像对对应的预设匹配标签作为深度学习模型的预期输出，对深度学习模型进行有监督训练。

预设匹配标签是匹配图像对对应的标签，表示匹配图像对中的照片对应的对象是匹配的。将各个匹配图像对，如 (A_1,A_2)，(B_1,B_2)，(C_1,C_2)，(D_1,D_2) 分别作为输入，将预设匹配标签，如"1"作为输出，进行有监督训练。

步骤 S6″，将各个非匹配图像对作为深度学习模型的输入，将非匹配图像对对应的预设非匹配标签作为深度学习模型的预期输出，对深度学习模型进行有监督训练。

预设非匹配标签是非匹配图像对对应的标签，表示非匹配图像对中的照片对应的对象是非匹配的。将各个非匹配图像对，如 (A_1,B_2)，(A_1,C_2)，(A_1,D_2)，(B_1,A_2)，(B_1,C_2)，(B_1,D_2)，(C_1,A_2)，(C_1,B_2)，(C_1,D_2)，(D_1,A_2)，(D_1,B_2)，(D_1,C_2) 分别作为输入，将预设匹配标签，如"0"作为输出，进行有监督训练。

进行自顶向下的监督训练，这一步是在第一步学习获得各层参数进的基础上，在最顶的编码层添加一个分类器，例如 Logistic 回归、SVM 等，而后通过带标签数据的监督训练，利用梯度下降法去微调整个网络参数。深度学习的第一步实质上是一个网络参数初始化过程，区别于传统神经网络初值随机初始化，深度学习模型是通过无监督训练输入数据的结构得到的，因而这个初值更接近全局最优，从而能够取得更好的效果。

步骤 S7″,得到已训练的深度学习模型。

经过上面的无监督训练和有监督训练后,得到已训练的深度学习模型。

在一个技术方案中,步骤 S4 包括:如果匹配结果与预设匹配标签相同,则确定第一对象与第二对象的匹配度为匹配成功;如果匹配结果与预设非匹配标签相同,则确定第一对象与第二对象的匹配度为匹配失败。

根据已训练的深度学习模型输出的匹配结果与预设匹配标签是否相同,来判断第一对象与第二对象是否匹配成功,可直接得出匹配结果,简单方便。预设匹配标签和预设非匹配标签的格式可根据需要自定义,如用标志位表示,如“1”表示预设匹配标签,“0”表示预设非匹配标签。

在一个技术方案中,步骤 S1 包括:接受请求匹配对象用户终端发送的匹配请求,匹配请求包括请求匹配对象图像数据,将请求匹配对象图像数据作为第一对象图像数据,获取待推荐对象图像数据作为第二对象图像数据。

请求匹配对象是指需要进行匹配的对象,请求匹配对象图像数据是请求匹配对象的图片或根据图片提取的特征数据,待推荐对象是指与请求匹配对象进行匹配的候选对象。匹配请求中可以直接携带待推荐对象图像数据,也可以从服务器获取待推荐对象图像数据。待推荐对象图像数据是与请求匹配对象进行匹配的候选对象的图片或根据图片提取的特征数据。待推荐对象图像数据可以根据待推荐对象的更新而动态的更新。

在一个技术方案中,从婚恋匹配网站获取注册用户的资料图片形成待推荐对象图像数据,可根据请求匹配对象的基本资料筛选待推荐对象。

步骤 S4 之后,进行以下步骤:如果匹配度为匹配成功,则向请求匹配对象用户终端发送待推荐对象对应的用户信息。

如果匹配度为匹配成功,则说明请求匹配对象与待推荐对象匹配,将待推荐对象对应的用户信息推送至请求匹配对象用户终端。只有匹配成功,才会进行推送,提高了请求匹配对象获取待推荐对象的有效性与客观性,提高请求匹配对象的感兴趣度,提高匹配效率,避免大量无效的待推荐对象的推送对用户形成干扰。

5.3.2 基于大数据与深度学习的婚恋对象匹配推荐系统

1. 基于大数据与深度学习的婚恋对象匹配推荐系统的基本模块

在一个技术方案中,如图 5-17 所示,提供了一种推送待推荐对象信息装置,包括:

(1) 获取模块 1,用于获取待匹配的第一和第二对象图像数据。

(2) 匹配结果获取模块 2,用于将第一和第二对象图像数据输入已训练的深度学习模型,深度学习模型的参数是根据已婚匹配对象对应的匹配图像大数据训练得到,获取已训练的深度学习模型输出的对应的匹配结果。

(3) 匹配度确定模块 3,用于根据匹配结果确定第一对象与第二对象的匹配度。

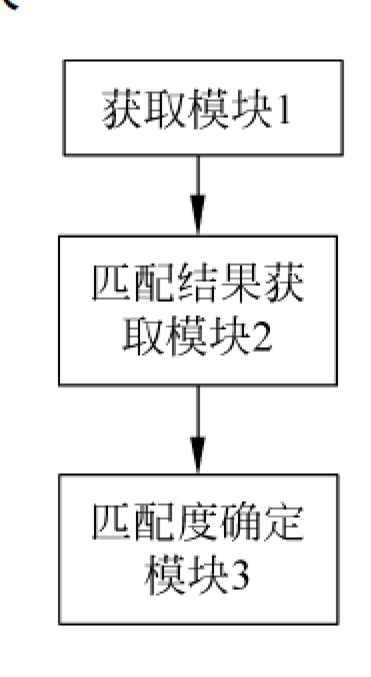


图 5-17 婚恋对象匹配数据处理装置的结构框图

2．基于大数据与深度学习的婚恋对象匹配推荐系统的扩展模块

在一个技术方案中，如图 5-18 所示，装置包括：

初始化配置模块 4，用于获取待匹配的对象对应的预设数据格式和匹配结果预设格式，将预设数据格式作为待输入格式，将匹配结果预设格式作为待输出格式，获取与待输入格式和待输出格式匹配的深度学习模型对应的初始配置信息，根据初始配置信息对深度学习模型进行初始化配置。

在一个技术方案中，初始化配置模块 4 还用于获取深度学习模型对应的候选输入格式和候选输出格式，获取输入权值系数和输出权值系数，计算待输入格式和候选输入格式的输入匹配度，计算待输出格式和候选输出格式的输出匹配度，根据输入权值系数、输入匹配度、输出权值系数、输出匹配度计算得到初始配置信息匹配度，根据初始配置信息匹配度获取与待输入格式和待输出格式最匹配的目标候选输入格式和目标候选输出格式，根据目标候选输入格式和目标候选输出格式获取对应的目标初始配置信息。

在一个技术方案中，如图 5-19 所示，装置包括：

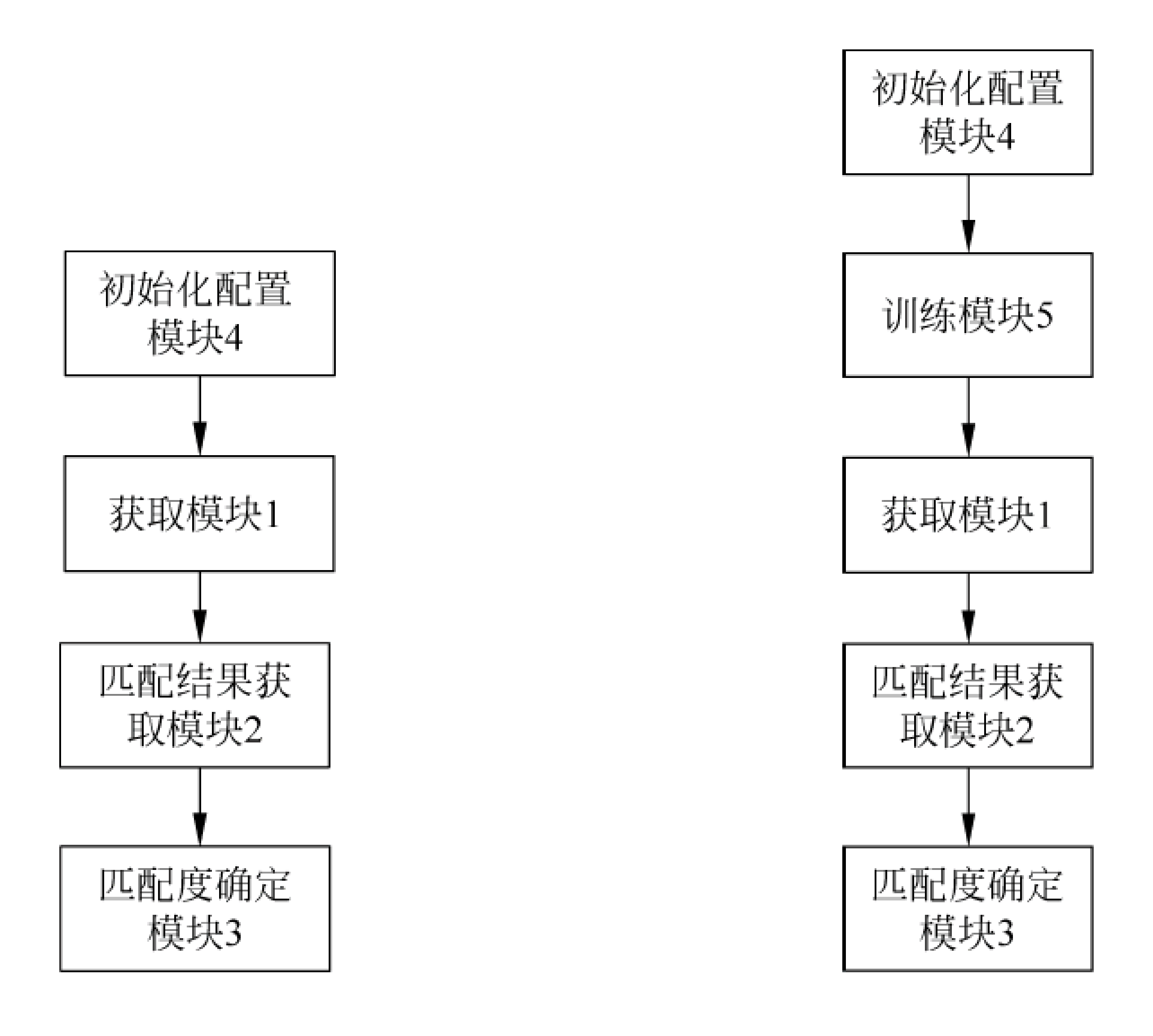


图 5-18　婚恋对象匹配数据处理装置的结构框图(一)

图 5-19　婚恋对象匹配数据处理装置的结构框图(二)

训练模块 5，用于获取训练图像大数据，包括已婚匹配对象对应的匹配图像大数据，从匹配图像大数据中提取对应的各个已婚匹配对象对应的匹配图像对，各个已婚匹配对象对应的匹配图像对组成匹配图像对集合，将任意两个已婚匹配图像对进行拆分组合形成非匹配图像对，各个非匹配图像对组成非匹配图像对集合，将各个匹配图像对和各个非匹配图像对分别作为深度学习模型的输入，对深度学习模型进行无监督训练，将各个匹配图像对作为深度学习模型的输入，将匹配图像对对应的预设匹配标签作为深度学习模型的预期输出，对深度学习模型进行有监督训练，将各个非匹配图像对作为深度学习模型的输入，将非匹配图

像对对应的预设非匹配标签作为深度学习模型的预期输出，对深度学习模型进行有监督训练，得到已训练的深度学习模型。

在一个技术方案中，匹配度确定模块 3 还用于如果匹配结果与预设匹配标签相同，则确定第一对象与第二对象的匹配度为匹配成功，如果匹配结果与预设非匹配标签相同，则确定第一对象与第二对象的匹配度为匹配失败。

在一个技术方案中，获取模块 1 还用于接受请求匹配对象用户终端发送的匹配请求，匹配请求包括请求匹配对象图像数据，将请求匹配对象图像数据作为第一对象图像数据，获取待推荐对象图像数据作为第二对象图像数据。

如图 5-20 所示，装置包括：

发送模块 6，用于如果匹配度为匹配成功，则向请求匹配对象用户终端发送待推荐对象对应的用户信息。

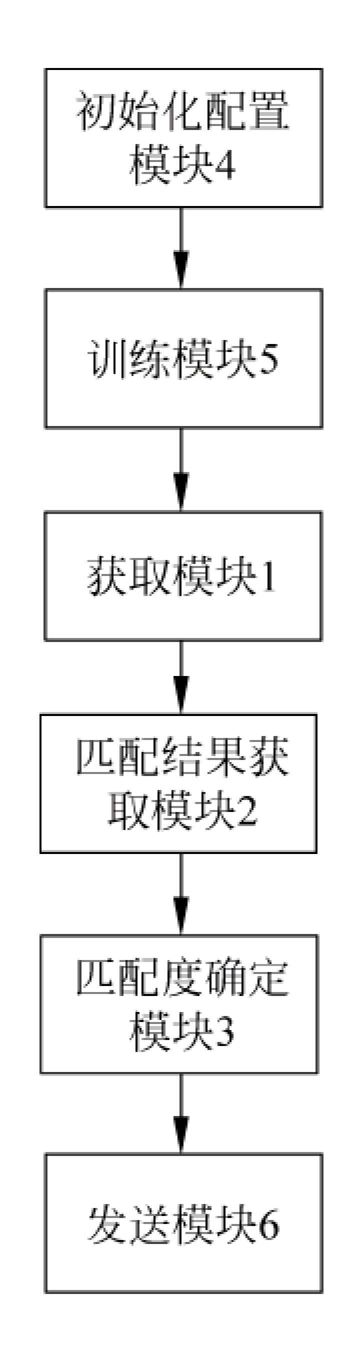


图 5-20 婚恋对象匹配数据处理装置的结构框图(三)



深度学习模型的检测诊断大数据应用

深度学习模型的应用很多，其中深度学习模型的检测诊断大数据有着非常重要的应用价值和前景。本章给出了两种深度学习模型的检测诊断大数据应用：基于深度学习模型和大数据的停车位检测、基于深度学习模型面向自动诊断的医疗数据处理。

6.1 基于深度学习模型和大数据的停车位检测

传统技术在进行停车位检测时，采用的是阈值法或模式识别法等方法[25,26]。这两种方法抗干扰能力差，而停车场中车来车往，不同车主停车模式千差万别，不同停车位所受的干扰不同，所以这种复杂的环境下会出现千变万化的检测数据，因此使用固定的几个阈值或固定的几个模式来识别停车位的相关状态，容易出现误识别，无法达到令人满意的检测精度。

本技术研究并给出了一种新的基于深度学习的停车位检测方法和系统。方法包括：获取停车场中停车位上有否驻车的相关检测数据；获取停车场中停车位上有否驻车的相关状态；将相关检测数据作为深度学习模型的输入，将相关状态作为深度学习模型的输出，对深度学习模型进行训练；获取当前时刻下停车场中待测停车位上有否驻车的相关检测数据；将相关检测数据输入训练后的深度学习模型进行深度学习；将深度学习模型的相应输出作为待测停车位上有否驻车的机器判定相关状态。本技术采用深度学习和深度学习模型对停车位的相关检测数据进行识别，可以从千变万化、干扰众多的停车位检测数据中去伪存真，得到的停车位状态能达到令用户满意的检测精度。

6.1.1 基于深度学习模型和大数据的停车位检测方法

1. 基于深度学习模型和大数据的停车位检测的基本方法

图 6-1 所示为一技术方案中停车位检测方法的流程图,包括下列步骤:

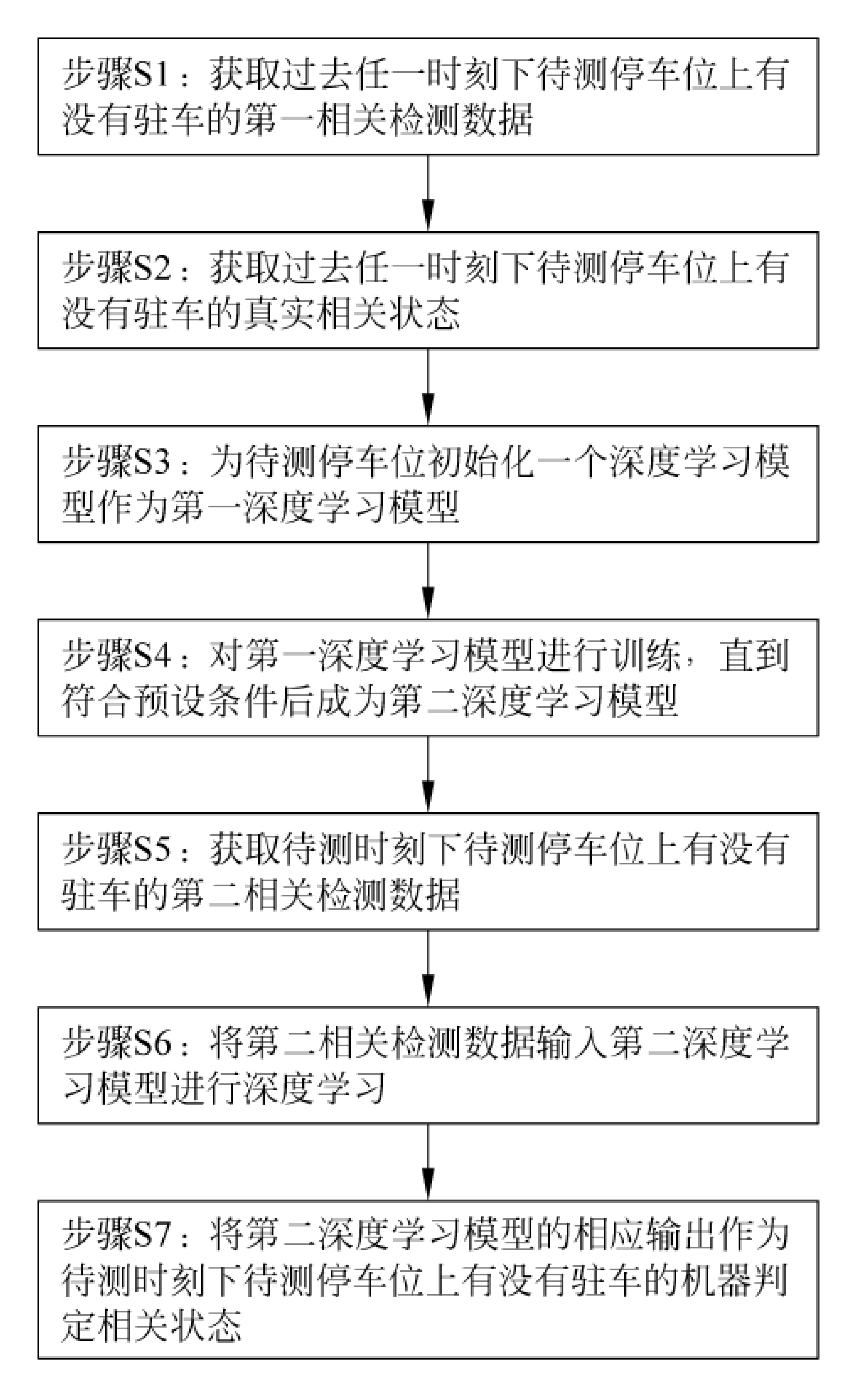


图 6-1 停车位检测方法的流程图

(1) 步骤 S1,获取过去任一时刻下待测停车位上有否驻车的第一相关检测数据。

停车场中每一个停车位上会设有采集用于判定该停车位上是否有驻车的数据的相应装置,例如地磁传感器、红外传感器、摄像头等,后续通过这些数据检测该停车位是否驻车。

(2) 步骤 S2,获取过去任一时刻下待测停车位上有否驻车的真实相关状态。

相关状态包括该停车位被占用和该停车位空闲。由于真实相关状态需要作为深度学习模型训练时的输入,因此要尽量保证状态准确。真实相关状态可以通过人工查看获取,也可以通过设备自动检测获取,但如前述,不管通过什么方式获取都要尽量保证状态准确。在其中一个技术方案中,停车位被占用和停车位空闲可以分别用 0、1 来表示,也可以使用其他预设的数值表示。

(3) 步骤 S3,为待测停车位初始化一个深度学习模型作为第一深度学习模型。

(4) 步骤 S4,对第一深度学习模型进行训练。

将第一相关检测数据输入第一深度学习模型的输入层、将真实相关状态输入第一深度

学习模型的输出层，对第一深度学习模型进行训练。一个停车位的第一相关检测数据和真实相关状态是一一对应的，即要在同一时刻获取该停车位的第一相关检测数据和真实相关状态。步骤 S1～S4 为深度学习模型的训练阶段，重复对第一深度学习模型进行多次训练，直到第一深度学习模型符合预设条件，将符合预设条件的第一深度学习模型作为第二深度学习模型，进行停车位检测。从步骤 S5 开始为检测阶段。

（5）步骤 S5，获取待测时刻下待测停车位上有否驻车的第二相关检测数据。采用与步骤 S1 中同样的装置获取该数据。

（6）步骤 S6，将第二相关检测数据输入第二深度学习模型进行深度学习。

通过第二深度学习模型进行停车位检测。

（7）步骤 S7，将第二深度学习模型的相应输出作为待测时刻下待测停车位上有否驻车的机器判定相关状态。

相关状态包括该停车位被占用和该停车位空闲。通过机器对待测停车位上有否驻车进行自动检测。

传统技术在停车位检测时，采用阈值法或模式识别法对停车位的检测数据进行识别，得到停车位的相关状态。但由于阈值和模式是预设好后就固定不变的，所以对千变万化、干扰众多的停车位检测数据进行识别，得到的停车位状态无法达到令用户满意的检测精度。

上述停车位检测方法，采用深度学习模型对停车位的相关检测数据进行识别，得到停车位的相关状态。由于深度学习模型的检测精度远远高于阈值法和模式识别法，所以可以从千变万化、干扰众多的停车位检测数据中去伪存真，得到的停车位状态能达到令用户满意的检测精度。

2. 基于深度学习模型和大数据的停车位检测的详细方法

在其中一个技术方案中，步骤 S3 具体包括将第一深度学习模型的输入层的数据结构初始化为待测停车位上有否驻车的第一相关检测数据的数据结构或可以转化成的数据结构，将第一深度学习模型的输出层的数据结构初始化为待测停车位上有否驻车的真实相关状态的数据结构或可以转化成的数据结构，并将第一深度学习模型初始化出预设层数个中间层，预设层数是大于或者等于 0 的整数。

步骤 S4 包括：将过去任一时刻下第一相关检测数据压缩成预设层数组的分辨率递减的中间数据，每一组中间数据与一个中间层对应；然后将每一组中间数据输入对应的中间层。

例如，预设层数为 3。将第一相关检测数据的分辨率压缩 50%，得到第一组中间数据；将第一组中间数据的分辨率压缩 50%，得到第二组中间数据；将第二组中间数据的分辨率压缩 50%，得到第三组中间数据。将第一组中间数据输入第一深度学习模型的第一个中间层，将第二组中间数据输入第一深度学习模型的第二个中间层，将第三组中间数据输入第一深度学习模型的第三个中间层。

在其中一个技术方案中，步骤 S4 判断第一深度学习模型是否符合预设条件，具体是通过如下步骤进行：获取某一时刻（例如距离待测时间最近的过去时刻）下待测停车位的第一相关检测数据及真实相关状态。将该时刻下待测停车位的第一相关检测数据输入第一深度学习模型的输入层，然后通过第一深度学习模型的深度学习得到第一深度学习模型输出层的输出结果。将该输出结果与该时刻下待测停车位的真实相关状态进行对比，如果对比得

到的误差不大于预设阈值，则符合预设条件，结束训练并将第一深度学习模型作为第二深度学习模型，进入检测阶段；否则判定为不满足预设条件，对第一深度学习模型进行下一次训练。

在其中一个技术方案中，步骤 S4 的训练是对每一个停车位都单独训练一个深度学习模型，即如果停车场有 N 个停车位，则训练出 N 个相应的深度学习模型。具体地，步骤 S1、步骤 S2 要获取某个停车位的很多个不同时间的第一相关检测数据和真实相关状态，然后在步骤 S4 中将每一对第一相关检测数据和真实相关状态分别作为深度学习模型的输入和输出进行训练。如果停车场有 N 个停车位，步骤 S1 和 S2 中就要将每个停车位的数据单独分组(即分成 N 组)，步骤 S4 训练时单独使用每个停车位的数据训练出一个对应停车位的深度学习模型。这样一来，不同的待检测停车位有不同的深度学习模型，可以更个性化地深度学习处于不同干扰下的不同的待检测停车位，能使得检测精度更高。

步骤 S4 对待测停车位的深度学习模型进行训练时，输入的第一相关检测数据除了该停车位有否驻车的检测数据以外，还需要输入该车位的相邻停车位有否驻车的检测数据。例如，要将该停车位左边的相邻停车位和右边的相邻停车位有否驻车的检测数据也作为该停车位的深度学习模型的输入。这样做的好处在于能够将相邻停车位对该停车位的干扰考虑进去(更远的停车位对该停车位的干扰相对较小，可以忽略)，从而使得训练出来的深度学习模型可以把相邻停车位的干扰一并考虑，使得相邻停车位的干扰不会影响深度学习模型对该停车位的检测精度。同样，步骤 S5 中的第二相关检测数据除了待测停车位有否驻车的检测数据以外，包括待测车位的相邻停车位的有否驻车的检测数据。可以理解的，在其他技术方案中也可以仅是将该停车位有否驻车的检测数据作为深度学习模型训练时的输入，而不输入相邻停车位有否驻车的检测数据，即第一相关检测数据不包括相邻停车位有否驻车的检测数据。

在另一个技术方案中，步骤 S4 的训练是面向停车场中的所有停车位进行训练，得到一个反映全部停车位的深度学习模型，这个深度学习模型可以供所有的待测停车位使用，这样可以节省训练所需的时间/计算资源。

步骤 S4 对待测停车位的深度学习模型进行训练时，输入的第一相关检测数据除了该停车位有否驻车的检测数据以外，还需要输入停车场中所有其他停车位上有否驻车的检测数据。同样，步骤 S5 中的第二相关检测数据除了待测停车位有否驻车的检测数据以外，包括停车场中所有其他停车位上有否驻车的检测数据。

步骤 S2 的相关状态和步骤 S7 的相关状态包括停车位上驻车车辆的类型，具体是根据车辆的大小进行划分的。例如，可以划分为小型车、中型车及大型车，分别用 2、3、4 表示，也可以使用其他预设的数值表示。划分的标准可以是车身的长度，例如设定两个长度阈值以划分出 3 种大小的车型。划分的标准可以采用本领域常用的划分标准，例如轿车的划分标准等，也可以根据停车位实际的尺寸自定义。

在其中一个技术方案中，第一和第二相关检测数据以图像来表示，深度学习模型的训练和深度学习采用用于图像识别的深度学习模型的训练和深度学习方法，以相应地对这些数据进行处理、完成对停车位的检测。

在其中一个技术方案中，上述停车位检测方法基于大数据进行应用。具体是步骤 S1 中将获取到的第一相关检测数据存储于大数据，步骤 S2 中将获取到的真实相关状态存储于大

数据，步骤 S4 中从大数据中读取第一相关检测数据作为深度学习模型的输入，步骤 S5 中将获取到的第二相关检测数据存储于大数据，步骤 S6 中从大数据中读取第二相关检测数据输入训练后的深度学习模型，步骤 S7 中将机器判定相关状态存储于大数据中。

6.1.2　基于深度学习模型和大数据的停车位检测系统

基于深度学习模型和大数据的停车位检测系统，包括：

（1）训练输入数据获取模块 1，用于获取过去任一时刻下停车场中的待测停车位上有否驻车的第一相关检测数据。

（2）训练输出数据获取模块 2，用于获取过去任一时刻下待测停车位上有否驻车的真实相关状态。

（3）初始化模块 3，用于为待测停车位初始化一个深度学习模型，作为第一深度学习模型。

（4）训练模块 4，将过去任一时刻下第一相关检测数据和真实相关状态分别输入第一深度学习模型的输入层和输出层，对第一深度学习模型进行训练；多次训练直到第一深度学习模型符合预设条件，将符合预设条件的第一深度学习模型作为第二深度学习模型。

（5）检测数据获取模块 5，用于获取待测时刻下待测停车位上有否驻车的第二相关检测数据。

（6）停车检测模块 6，用于将第二相关检测数据输入第二深度学习模型进行深度学习。

（7）停车判定模块 7，用于将第二深度学习模型的相应输出作为待测时刻下待测停车位上有否驻车的机器判定相关状态。

6.2　基于深度学习模型面向自动诊断的医疗数据处理

随着计算机技术的发展，医院利用计算机来记录患者的医疗数据，并将医疗数据存储到服务器。每天有大量的患者到医院就诊，医生将每个患者在就诊过程中产生的医疗数据存储到服务器中。

然而，在将医疗数据存储到服务器中后，大量的医疗数据需要占用较多的存储空间，并且这些数据没有得到充分的利用。同时，患者到医院就诊时，需要耗费大量时间进行各种类型的身体检查，医生需要耗费时间对各种检查结果进行详细研究，才能对患者进行确诊从而得到确诊结果，患者的确诊结果的生成效率较低[27,28]。

本技术研究并给出了一种新的医疗数据处理系统和装置，该系统包括体检数据采集设备和数据服务器；体检数据采集设备用于获取覆盖待确诊患者身体各部分的第一体检数据，并将第一体检数据发送给数据服务器；数据服务器利用深度学习模型对第一体检数据进行运算得到第一体检数据对应的确诊结果；深度学习模型是根据覆盖已确诊患者身体各部分的第二体检数据和第二体检数据对应的确诊数据训练得到的。本技术提供的医疗数据处理系统和装置，在确诊结果的生成过程中，无须耗费过多时间进行各项身体检查，也无须

医生对各项检查结果耗费时间进行详细研究，只需在接收到体检数据采集设备发送的第一体检数据后，即可根据学习模型进行运算得到确诊结果，提高了确诊结果的生成效率。

6.2.1 基于深度学习模型面向自动诊断的医疗数据处理系统

1. 基于深度学习模型面向自动诊断的医疗数据处理系统的应用环境

图 6-2 所示为医疗数据处理系统的应用环境图，医疗数据分析系统包括体检数据采集设备 1 和数据服务器 2，体检数据采集设备 1 和数据服务器 2 通过网络连接。体检采集设备 1 包括采集控制终端 1-1、扫描部件 1-2 和采集平台 1-3，采集控制终端 1-1 通过控制扫描部件 1-2 和采集平台 1-3 采集位于采集平台 1-3 上患者的体检数据。扫描部件 1-2 具体可以是 B 超设备中的超声扫描仪，还可以是 CT 设备中 X 线管和探测器，也可以是 MRI (Magnetic Resonance Imaging，磁共振成像)成像设备中的磁场发生器、无线电波发生器和探测器。

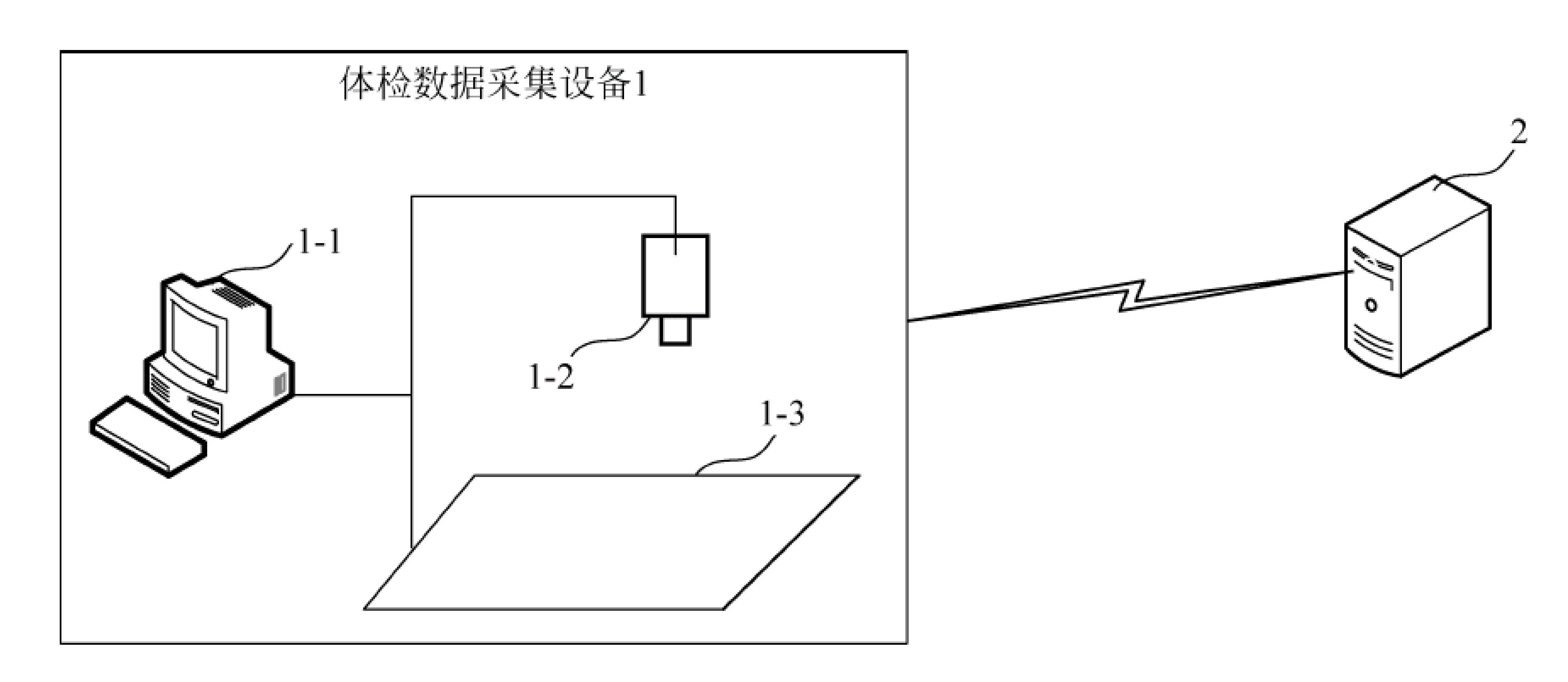


图 6-2　医疗数据处理系统的应用环境图

在一个技术方案中，体检数据采集设备 1 用于获取覆盖待确诊患者身体各部分的第一体检数据，并将第一体检数据发送给数据服务器 2。

体检数据采集设备 1 对待确诊患者身体各部位进行检查，通过对待确诊患者身体各部分的检查生成患者的第一体检数据，第一体检数据为覆盖待确诊患者身体各部分的体检数据，包括对待确诊患者各部分进行检查得到的数据，具体可以包括待确诊患者的全身体检数据。体检数据采集设备 1 在获取到第一体检数据后，将其通过网络发送给数据服务器 2。

在一个技术方案中，体检数据采集设备 1 可以是医疗扫描设备，医疗扫描设备用于对待确诊患者身体部分进行扫描来获取待确诊患者的第一体检数据。

数据服务器 2 用于利用深度学习模型对第一体检数据进行运算得到第一体检数据对应的确诊结果；深度学习模型是根据覆盖已确诊患者身体各部分的第二体检数据及其对应的确诊数据训练得到的。

数据服务器 2 获取已经确诊患者的第二体检数据及其对应的确诊数据，以获取到的第二体检数据及其对应的确诊数据作为训练样本，以第二体检数据作为深度学习模型的输入，并将第二体检数据对应的已确诊患者的确诊数据作为深度学习模型的输出，通过训练得到

深度学习模型。在数据服务器2将获取到的第一体检数据输入到训练得到的深度学习模型中,深度学习模型对第一体检数据进行运算得到对应的确诊结果,其中确诊结果可以疾病名称,也可以是疾病名称对应的编码。

本技术方案中,数据服务器接收体检数据采集设备采集的待确诊患者的第一体检数据,第一体检数据覆盖待确诊患者身体各部分,可以全面反映患者身体的健康状况。利用覆盖已确诊患者身体各部分的第二体检数据训练得到的深度学习模型对覆盖待确诊患者身体各部分的第一体检数据进行运算,通过深度学习模型的运算得到第一体检数据对应的确诊结果。在确诊结果的生成过程中,无须耗费过多时间进行各项身体检查,也无须医生对各项检查结果耗费时间进行详细研究,只需在接收到体检数据采集设备发送的第一体检数据后,即可根据深度学习模型进行运算得到确诊结果,提高了确诊结果的生成效率,进而辅助医生更加快速、准确、合理地做出诊断。

在一个技术方案中,体检数据采集设备1还用于获取第二体检数据,并将其发送给数据服务器。

体检数据采集设备1为医疗扫描设备,医疗扫描设备用于对已确诊患者身体部分进行扫描来获取已确诊患者的第二体检数据。体检数据采集设备1对已确诊患者身体各部分进行扫描得到已确诊患者的医疗图像,提取医疗图像中的特征区域进行分析得到已确诊患者的覆盖身体各部分的第二体检数据。医疗图像具体可以是CT图像、B超图像和MRI图像中的至少一种。体检数据采集设备1将第二体检数据发送给数据服务器2。

数据服务器还用于将第二体检数据相应的已确诊患者的确诊数据进行关联存储到医疗数据库中;从医疗数据库中提取第二体检数据及其对应的确诊数据;根据提取到的第二体检数据及其对应的确诊数据训练得到深度学习模型。

数据服务器2在接收到第二体检数据后,提取第二体检数据中已经确诊患者的患者标识,再从确诊数据库中提取患者标识对应的确诊数据。数据服务器2将同一个患者标识对应的确诊数据和第二体检数据进行关联并存储到医疗数据库中。

在一个技术方案中,数据服务器2同一个患者标识的确诊数据和第二体检数据进行关联存储,具体可以将同一个患者标识对应的确诊数据和第二体检数据对应存储在医疗数据库中。将同一患者标识的确诊数据和第二体检数据进行关联存储,还可以分别提取同一个患者标识的确诊数据和第二体检数据中的特征数据,建立特征数据的对应关系表,通过特征数据的对应关系表将同一个患者标识对应的确诊数据和第二体检数据关联,并将关联后的确诊数据和第二体检数据存储到医疗数据库中。也可以通过患者标识将同一个患者标识对应的确诊数据和第二体检数据进行关联,将患者标识、确诊数据和第二体检数据对应存储在医疗数据库中。

数据服务器2从医疗数据库中提取多个已确诊患者的患者标识对应的第二体检数据及其对应的确诊数据,并根据提取到的第二体检数据及其对应的确诊数据训练深度学习模型。

本技术方案中,将已确诊患者的确诊数据和第二体检数据存储在医疗数据库中,使医疗数据库中存储着海量已确诊患者的确诊数据和第二体检数据,并且同一患者标识对应的确诊数据和体检数据关联存储在医疗数据库中,为训练深度学习模型提供了大数据支撑,通过医疗数据库中海量已确诊患者的数据对深度学习模型进行训练,提高了通过深度学习模型生成确诊结果的准确性。

在一个技术方案中，数据服务器 2 还用于根据确诊结果确定异常体检指标；在第一体检数据中提取异常体检指标对应的数值；根据异常体检指标对应的数值和确诊结果生成确诊报告。

体检指标是在体检数据中能够反映人身体健康状况的检测项目，例如心率、血压和尿酸等，每个体检指标对应的数值都有一个正常范围，根据体检指标的数值是否在正常范围内可以判断人身体的健康状况。体检指标的数值超出该体检指标对应的正常范围，则该体检指标为异常体检指标。

数据服务器 2 在生成确诊结果后，根据确诊结果与异常体检指标的对应关系确定异常体检指标的体检指标标识，根据确定的体检指标标识在第一体检指标中查找异常体检指标的数值并提取，根据提取到的异常体检指标的数值和确诊结果生成相应的待确诊患者的确诊报告。其中确诊结果与异常体检指标的对应关系可以通过数据服务器 2 以医疗数据库中存储的部分或全部数据作为样本进行分析获得。

本技术方案中，在生成确诊结果后，根据确诊结果提取第一体检数据中的异常体检指标的数值，根据确诊结果和异常体检指标的数值生成确诊报告，确诊报告可以准确详细地反映待确诊患者的健康状况。

在一个技术方案中，数据服务器还用于提取第一体检数据中各体检指标的数值，分别对各体检指标的数值进行编码生成各体检指标对应的编码数据，编码数据包括指标标识和指标数值；将编码数据输入深度学习模型，使深度学习模型将编码数据输入到与编码数据中指标标识对应的节点进行运算得到运算结果；根据运算结果确定第一体检数据对应的疾病名称。

数据服务器 2 提取第一体检数据中各体检指标的数值，查找与体检指标名称匹配的指标标识，指标标识用于区别不同体检指标，具体可以是字符和数字中的至少一种。将指标标识添加到体检指标的数值中形成编码数据，编码数据中包括指标标识和指标数值。数据服务器 2 将编码数据输入到深度学习模型中，深度学习模型读取编码数据中的指标标识，并查找指标标识对应的节点，提取编码数据中的指标数值，将指标数值输入指标标识对应的节点。深度学习模型根据输入的指标数值运算得到运算结果，运算结果可以是字符或数值中的至少一种，查找与运算结果对应的疾病名称，并将查找到的疾病名称输出作为待确诊患者的确诊结果。

本技术方案中，通过各体检指标进行编码生成编码数据，根据编码数据中指标标识将编码数据中的指标数值输入到相应的深度学习模型节点，通过对体检指标进行编码保证了深度学习模型节点的输入准确率，从而使得运算结果更加准确。

在一个技术方案中，数据服务器 2 还用于提取第一体检数据中各体检指标的数值，对各体检指标的数值进行标准化处理得到各体检指标的标准化值；将各体检指标的标准化值输入深度学习模型，使深度学习模型将各体检指标的标准化值输入相应的节点，并根据节点之间的权重值对各体检指标的标准化值进行运算得到第一体检数据对应的确诊结果。

数据服务器 2 提取第一体检数据中各体检指标的数值，通过标准化处理将第一体检数据中各体检指标的数值，根据各体检指标的映射关系将各体检指标的数值映射到特定的数值区间中的数值，其中特定的数值区间具体可以是[−1,0]、[0,1]或[−1,1]。

在一个技术方案中，数据服务器 2 利用体检指标的数值减去该体检指标的最小值得到差 x_1，再利用该体检指标的最大值减去最小值得到差 x_2，利用 x_1 除以 x_2 得到体检指标的

数值在数值区间[0,1]中的标准化值。

在一个技术方案中，数据服务器 2 还用于提取第一体检数据中各体检指标对应的值，查找提取到的值不是数值的体检指标，将查找到的体检指标的值转化成数值，以各体检指标对应的数值为标准化值。例如，体检指标“尿葡萄糖”对应的值是“阴性”，该体检指标对应的值不是数值，将“阴性”转化为“1”，如果是“阳性”，则将“阳性”转化为“0”。

在一个技术方案中，第一体检指标中各体检指标对应的数值格式不同，数据服务器 2 对各体检指标对应的数值的格式进行调整，使各体检指标对应的数值格式一致，以格式调整后的数值作为各体检指标对应的标准化值。

数据服务器 2 在对各体检指标的数值进行标准化处理后得到各体检指标的标准化值，将各体指标的标准化值输入深度学习模型，深度学习模型将各体检指标的标准化值输入各体检指标对应的节点，深度学习模型根据各节点之间的权重值对输入的各体检指标的标准化值进行加权运算，通过多次加权运算得到最终的运算结果，根据运算结果查找对应的疾病名称，并将疾病名称作为第一体检数据对应的确诊结果。

本技术方案中，数据服服务器 2 将第一体检数据进行标准化处理，得到第一体检数据的各体检指标的标准化值，利用深度学习模型对各体检指标的标准化值进行运算，得到确诊结果。通过对标准化值进行计算，节省了确诊结果的生成过程中的运算资源，提高了运算效率，同时也提高了确诊结果的准确性和生成效率。

2. 基于深度学习模型面向自动诊断的医疗数据处理系统的模块构成

如图 6-3 所示，在一个技术方案中，提供一种医疗数据处理装置 2，该装置包括：第一数据获取模块 2-1 和确诊结果获得模块 2-2。

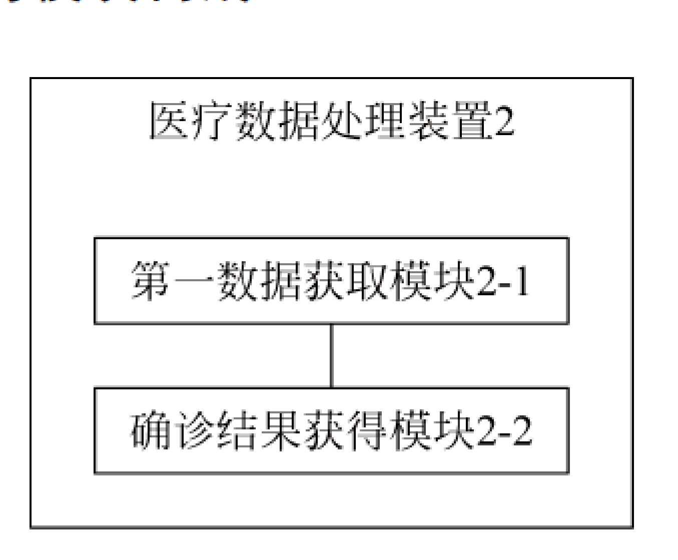


图 6-3　医疗数据处理装置的结构框图(一)

第一数据获取模块 2-1，用于获取覆盖待确诊患者身体各部分待确诊患者的第一体检数据。

对待确诊患者身体各部位将检查，通过对待确诊患者身体各部分的检查生成患者的第一体检数据，第一体检数据为覆盖待确诊患者身体各部分的体检数据，包括对待确诊患者各部分进行检查得到的数据，具体包括待确诊患者的全身体检数据。具体可以对待确诊患者进行覆盖身体各部分的扫描来获取待确诊患者的第一体检数据。

确诊结果获得模块 2-2，用于利用深度学习模型对第一体检数据进行运算得到第一体检数据对应的确诊结果；深度学习模型是根据覆盖已确诊患者身体各部分的第二体检数据及其对应的确诊数据训练得到的。

其中深度学习模型根据获取到的已经确诊患者的第二体检数据及其对应的确诊数据作为训练样本，以第二体检数据作为深度学习模型的输入，并以第二体检数据对应的已确诊患者的确诊数据作为深度学习模型的输出，通过训练得到的。将获取到的第一体检数据输入到训练得到的深度学习模型中，深度学习模型对第一体检数据进行运算得到第一体检数据对应的确诊结果，其中确诊结果可以疾病名称，也可以是疾病名称对应的编码。

本技术方案中，获取待确诊患者的第一体检数据，第一体检数据覆盖待确诊患者身体各部分，可以全面反映患者身体的健康状况。利用覆盖已确诊患者身体各部分的第二体检数

据训练得到的深度学习模型对覆盖待确诊患者身体各部分的第一体检数据进行运算，通过深度学习模型的运算得到第一体检数据对应的确诊结果。如图 6-4 所示，在一个技术方案中，该医疗数据处理装置 2 包括：第二数据获取模块 2-3、数据关联存储模块 2-4、关联数据提取模块 2-5 和深度学习模型训练模块 2-6。

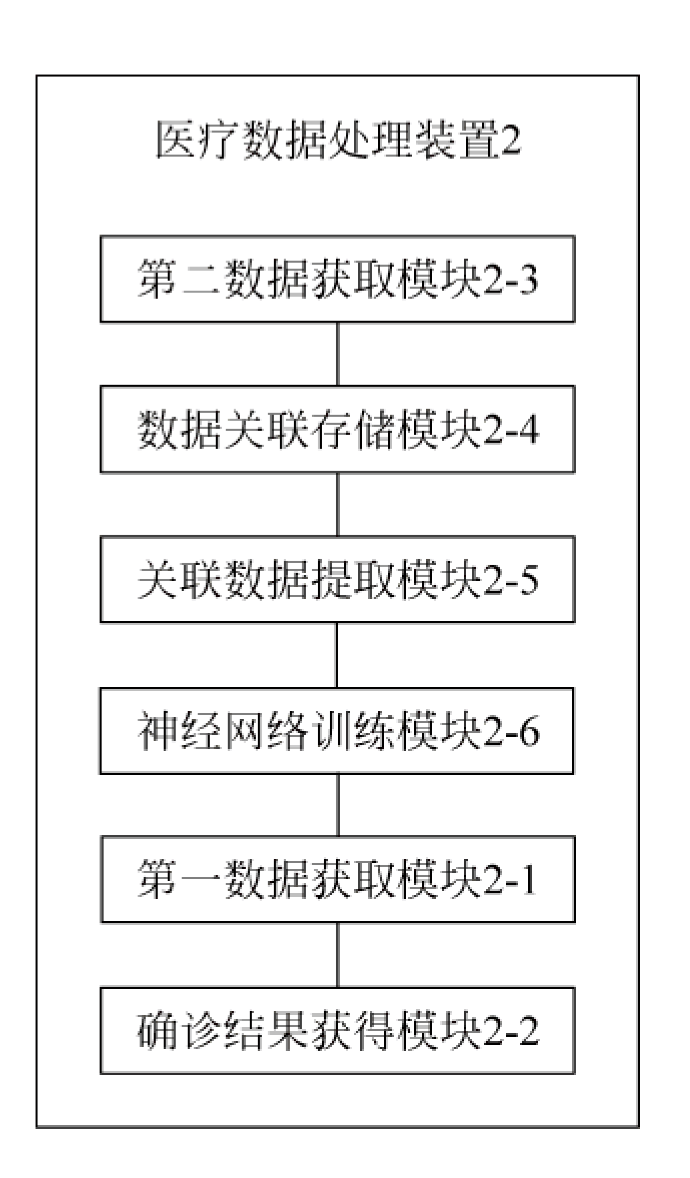


图 6-4　医疗数据处理装置的结构框图（二）

第二数据获取模块 2-3，用于获取第二体检数据。

利用医疗扫描设备对已确诊患者身体部分进行扫描来获取已确诊患者的第二体检数据。具体可以是对已确诊患者身体各部分进行扫描得到已确诊患者的医疗图像，提取医疗图像中的特征区域进行分析得到已确诊患者的覆盖身体各部分的第二体检数据。医疗图像具体可以是 CT 图像、B 超图像和 MRI 图像中的至少一种。体检数据采集设备 1 将第二体检数据发送给数据服务器 2。

数据关联存储模块 2-4，用于将第二体检数据相应的已确诊患者的确诊数据进行关联存储到医疗数据库中。

在获取到第二体检数据后，提取第二体检数据中已经确诊患者的患者标识，再从确诊数据库中提取患者标识对应的确诊数据。将同一个患者标识对应的确诊数据和第二体检数据进行关联并存储到医疗数据库中。

在一个技术方案中，数据服务器 2 同一个患者标识的确诊数据和第二体检数据进行关联存储，具体可以将同一个患者标识对应的确诊数据和第二体检数据对应存储在医疗数据库中。将同一患者标识的确诊数据和第二体检数据进行关联存储，还可以分别提取同一个患者标识的确诊数据和第二体检数据中的特征数据，建立特征数据的对应关系表，通过特征数据的对应关系表将同一个患者标识对应的确诊数据和第二体检数据关联，并将关联后的确诊数据和第二体检数据存储到医疗数据库中。也可以通过患者标识将同一个患者标识对应的确诊数据和第二体检数据进行关联，将患者标识、确诊数据和第二体检数据对应存储在医疗数据库中。

关联数据提取模块 2-5，用于从医疗数据库中提取第二体检数据及其对应的确诊数据。

深度学习模型训练模块 2-6，用于根据提取到的第二体检数据及其对应的确诊数据训练得到深度学习模型。

从医疗数据库中提取多个已确诊患者的患者标识对应的第二体检数据及其对应的确诊数据，以获取到的第二体检数据及其对应的确诊数据作为训练样本，以第二体检数据作为深度学习模型的输入，并第二体检数据对应的已确诊患者的确诊数据作为深度学习模型的输出，通过训练得到深度学习模型。

本技术方案中，将已确诊患者的确诊数据和第二体检数据存储在医疗数据库中，使医疗数据库中存储着海量已确诊患者的确诊数据和第二体检数据，并且同一患者标识对应的确诊数据和体检数据关联存储在医疗数据库中，为训练深度学习模型提供了大数据支撑，通过医疗数据库中海量已确诊患者的数据对深度学习模型进行训练，提高了通过深度学习模型生成确诊结果的准确性。

如图 6-5 所示，在一个技术方案中，该医疗数据处理装置 2 包括：异常指标确定模块 2-7、

异常数值提取模块 2-8 和确诊报告生成模块 2-9。

异常指标确定模块 2-7,用于根据确诊结果确定异常体检指标。

异常数值提取模块 2-8,用于在第一体检数据中提取异常体检指标对应的数值。

确诊报告生成模块 2-9,用于根据异常体检指标对应的数值和确诊结果生成确诊报告。

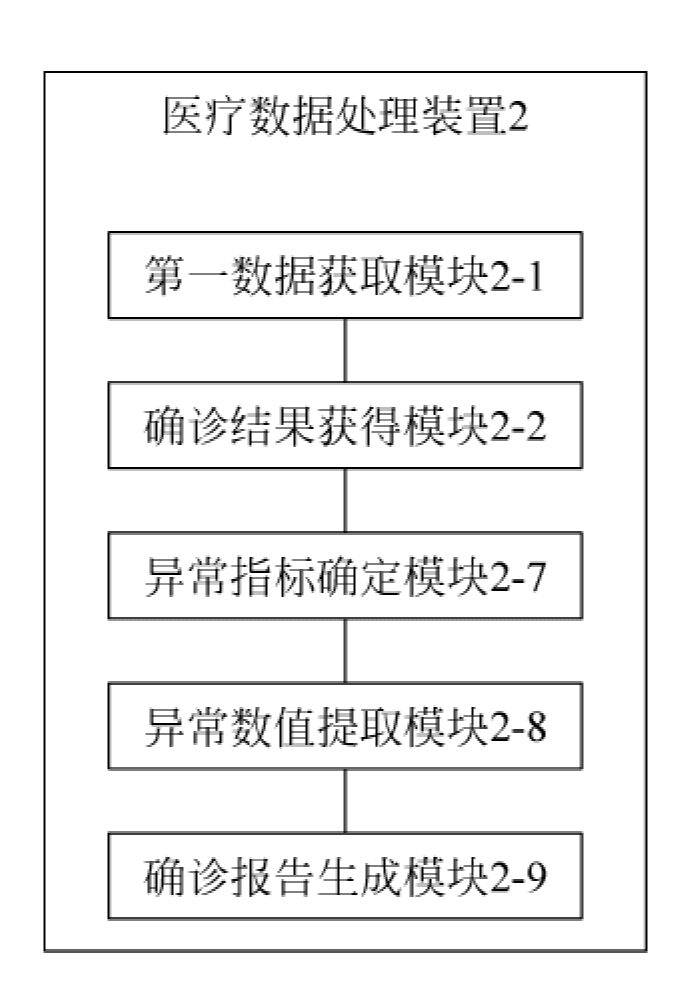


图 6-5　医疗数据处理装置的结构框图(三)

体检指标是在体检数据中能够反映人身体健康状况的检测项目,例如心率、血压和尿酸等,每个体检指标对应的数值都有一个正常范围,根据体检指标的数值是否在正常范围内可以判断人身体的健康状况。体检指标的数值超出该体检指标对应的正常范围,则该体检指标为异常体检指标。

在生成确诊结果后,根据确诊结果与异常体检指标的对应关系确定异常体检指标的体检指标标识,根据确定的体检指标标识在第一体检指标中查找异常体检指标的数值并提取,根据提取到的异常体检指标的数值和确诊结果生成相应的待确诊患者的确诊报告。其中确诊结果与异常体检指标的对应关系可以通过数据服务器 2 以医疗数据库中存储的部分或全部数据作为样本进行分析获得,也可以是预先设置的确诊结果与异常体检指标的对应关系表。

本技术方案中,在生成确诊结果后,根据确诊结果提取第一体检数据中的异常体检指标的数值,根据确诊结果和异常体检指标的数值生成确诊报告,确诊报告可以准确详细地反映待确诊患者的健康状况。

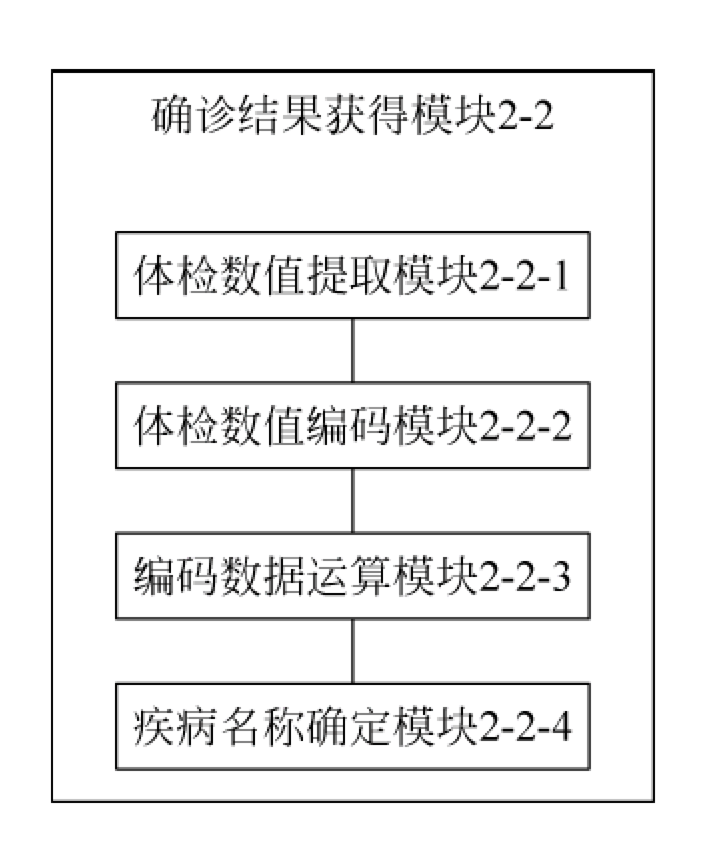


图 6-6　确诊结果获取模块的结构框图(一)

如图 6-6 所示,在一个技术方案中,确诊结果获得模块 2-2 包括:体检数值提取模块 2-2-1、体检数值编码模块 2-2-2、编码数据运算模块 2-2-3 和疾病名称确定模块 2-2-4。

体检数值提取模块 2-2-1,用于提取第一体检数据中各体检指标的数值。

体检数值编码模块 2-2-2,用于分别对各体检指标的数值进行编码生成各体检指标对应的编码数据,包括指标标识和指标数值。

编码数据运算模块 2-2-3,用于将编码数据输入深度学习模型,使深度学习模型将编码数据输入到与编码数据中指标标识对应的节点进行运算得到运算结果。

疾病名称确定模块 2-2-4,用于根据运算结果确定第一体检数据对应的疾病名称。

体检数据提取模块提取第一体检数据中各体检指标的数值,查找与体检指标名称匹配的指标标识,指标标识用于区别不同的体检指标,具体可以是字符和数字中的至少一种。将指标标识添加到体检指标的数值中形成编码数据,包括指标标识和指标数值,编码数据具体可以是二进制数。数据服务器 2 将编码数据输入到深度学习模型中,深度学习模型读取编码数据中的指标标识,并查找指标标识对应的节点,提取编码数据中的指标数值,将指标数

值输入指标标识对应的节点。深度学习模型根据输入的指标数值运算得到运算结果，运算结果可以是字符或数值中的至少一种，查找与运算结果对应的疾病名称，并将查找到的疾病名称输出作为待确诊患者的确诊结果。

本技术方案中，通过各体检指标进行编码生成编码数据，根据编码数据中指标标识将编码数据中的指标数值输入到相应的深度学习模型节点，通过对体检指标进行编码保证了深度学习模型节点的输入准确率，从而使得运算结果更加准确，进而辅助医生更加快速、准确、合理地做出诊断。

如图 6-7 所示，在一个技术方案中，确诊结果获得模块 2-2 包括：指标数值提取模块 2-2-5、指标数值处理模块 2-2-6 和标准化值运算模块 2-2-7。

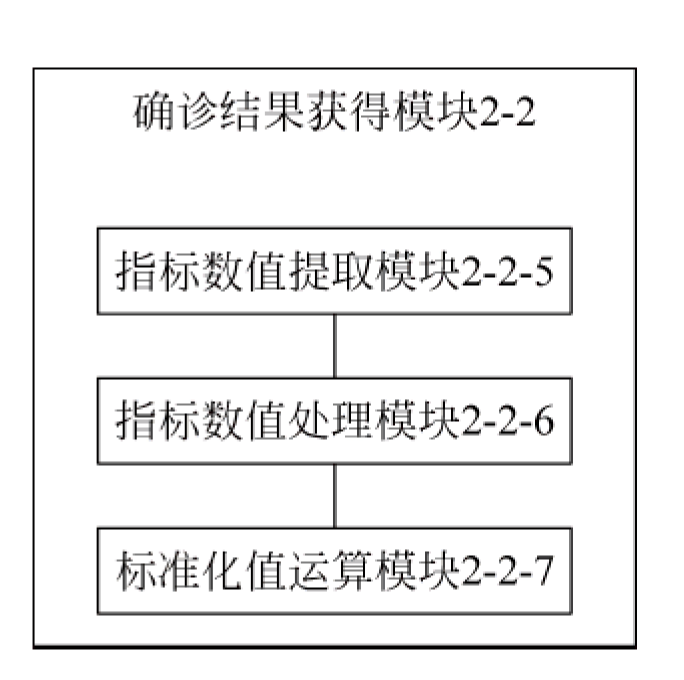


图 6-7　确诊结果获取模块的结构框图(二)

指标数值提取模块 2-2-5，用于提取第一体检数据中各体检指标的数值。

指标数值处理模块 2-2-6，用于对各体检指标的数值进行标准化处理得到各体检指标的标准化值。

提取第一体检数据中各体检指标的数值，通过标准化处理将第一体检数据中各体检指标的数值根据各体检指标的映射关系映射到特定的区间中的数值，其中特定的区间具体可以是[－1,0]、[0,1]或[－1,1]。

在一个技术方案中，利用体检指标的数值减去该体检指标的最小值得到差 x_1，再利用该体检指标的最大值减去最小值得到差 x_2，利用 x_1 除以 x_2 得到体检指标的数值在区间[0,1]中的标准化值。

在一个技术方案中，数据服务器 2 还用于提取第一体检数据中各体检指标对应的值，如查找提取到的值不是数值的体检指标，将查找到的体检指标的值转化成数值，以各体检指标对应的数值为标准化值。例如，体检指标“尿葡萄糖”对应的值是“阴性”，该体检指标对应的值不是数值，将“阴性”转化为“1”，如果是“阳性”，则将“阳性”转化为“0”。

在一个技术方案中，第一体检指标中各体检指标对应的数值格式不同，数据服务器 2 对各体检指标对应的数值的格式进行调整，使各体检指标对应的数值格式一致，以格式调整后的数值作为各体检指标对应的标准化值。

标准化值运算模块 2-2-7，用于将各体检指标的标准化值输入深度学习模型，使深度学习模型将各体检指标的标准化值输入相应的节点，并根据深度学习模型中节点之间的权重值对各体检指标的标准化值进行运算得到第一体检数据对应的确诊结果。

在对各体检指标的数值进行标准化处理后得到各体检指标的标准化值，将各体指标的标准化值输入深度学习模型，深度学习模型将各体检指标的标准化值输入各体检指标对应的节点，深度学习模型根据各节点之间的权重值对输入的各体检指标的标准化值进行加权运算，通过多次加权运算得到最终的运算结果，根据运算结果查找对应的疾病名称，并将疾病名称作为第一体检数据对应的确诊结果。

本技术方案中，数据服务器 2 将第一体检数据进行标准化处理，得到第一体检数据的各体检指标的标准化值，利用深度学习模型对各体检指标的标准化值进行运算，得到确诊结果。通过对标准化值进行计算，节省了确诊结果的生成过程中的运算资源，提高了运算效

率，同时也提高了确诊结果的准确性和生成效率。

6.2.2　基于深度学习模型面向自动诊断的医疗数据处理方法

如图 6-8 所示，在一个技术方案中，提供一种医疗数据处理方法，该方法具体包括以下步骤：

步骤S1：获取覆盖待确诊患者身体各部分的第一体检数据

↓

步骤S2：利用学习模型对第一体检数据进行运算得到第一体检数据对应的确诊结果；学习模型是根据覆盖已确诊患者身体各部分的第二体检数据及其对应的确诊数据训练得到的

图 6-8　医疗数据处理方法的流程示意图

（1）步骤 S1，获取覆盖待确诊患者身体各部分的第一体检数据。

（2）步骤 S2，利用深度学习模型对第一体检数据进行运算得到第一体检数据对应的确诊结果；深度学习模型是根据覆盖已确诊患者身体各部分的第二体检数据及其对应的确诊数据训练得到的。

本技术方案中，获取待确诊患者的第一体检数据，第一体检数据覆盖待确诊患者身体各部分，可以全面反映患者身体的健康状况。利用覆盖已确诊患者身体各部分的第二体检数据训练得到的深度学习模型对覆盖待确诊患者身体各部分的第一体检数据进行运算，通过深度学习模型的运算得到第一体检数据对应的确诊结果。

（3）在一个技术方案中，步骤 S1 之前包括训练深度学习模型的步骤，具体步骤如下：获取第二体检数据；将第二体检数据相应的已确诊患者的确诊数据进行关联存储到医疗数据库中；从医疗数据库中提取关联的第二体检数据及其对应的确诊数据；根据提取到的第二体检数据及其对应的确诊数据训练得到深度学习模型。

本技术方案中，将已确诊患者的确诊数据和第二体检数据存储在医疗数据库中，使医疗数据库中存储着海量已确诊患者的确诊数据和第二体检数据，并且同一患者标识对应的确诊数据和体检数据关联存储在医疗数据库中，为训练深度学习模型提供了大数据支撑，通过医疗数据库中海量已确诊患者的数据对深度学习模型进行训练，提高了通过深度学习模型生成确诊结果的准确性。

（4）在一个技术方案中，步骤 S2 之后包括确诊报告的生成步骤：根据确诊结果确定异常体检指标；在第一体检数据中提取异常体检指标对应的数值；根据异常体检指标对应的数值和确诊结果生成确诊报告。

本技术方案中，在生成确诊结果后，根据确诊结果提取第一体检数据中的异常体检指标的数值，根据确诊结果和异常体检指标的数值生成确诊报告，确诊报告可以准确详细地反映待确诊患者的健康状况，进而辅助医生更加快速、准确、合理地做出诊断。

（5）在一个技术方案中，步骤 S2 具体包括指标数值的编码步骤：提取第一体检数据中各体检指标的数值；分别对各体检指标的数值进行编码生成各体检指标对应的编码数据，

编码数据包括指标标识和指标数值；将编码数据输入深度学习模型，使深度学习模型将编码数据输入到与编码数据中指标标识对应的节点进行运算得到运算结果；根据运算结果确定第一体检数据对应的疾病名称。

本技术方案中，通过各体检指标进行编码生成编码数据，根据编码数据中指标标识将编码数据中的指标数值输入到相应的深度学习模型节点，通过对体检指标进行编码保证了深度学习模型节点的输入准确率，从而使得运算结果更加准确。

（6）在一个技术方案中，步骤 S2 具体包括标准化处理步骤：提取第一体检数据中各体检指标的数值；指标数值处理模块，用于对各体检指标的数值进行标准化处理得到各体检指标的标准化值；标准化值运算模块，用于将各体检指标的标准化值输入深度学习模型，使深度学习模型将各体检指标的标准化值输入相应的节点，并根据深度学习模型中节点之间的权重值对各体检指标的标准化值进行运算得到第一体检数据对应的确诊结果。

本技术方案中，数据服务器 2 将第一体检数据进行标准化处理，得到第一体检数据的各体检指标的标准化值，利用深度学习模型对各体检指标的标准化值进行运算，得到确诊结果。通过对标准化值进行计算，节省了确诊结果的生成过程中的运算资源，提高了运算效率，同时也提高了确诊结果的准确性和生成效率，进而辅助医生更加快速、准确、合理地做出诊断。

参 考 文 献

[1] Schmidhuber J. Deep learning in neural networks：An overview[J]. Neural networks，2015，61：85-117.

[2] Bengio Y，Goodfellow I J，Courville A. Deep learning[J]. Nature，2015，521：436-444.

[3] Pinheiro P，Collobert R. Recurrent convolutional neural networks for scene labeling[C]. International Conference on Machine Learning. 2014：82-90.

[4] Bar Y，Diamant I，Wolf L，et al. Chest pathology detection using deep learning with non-medical training[C]. Biomedical Imaging (ISBI)，2015 IEEE 12th International Symposium on. IEEE，2015：294-297.

[5] Esteva A，Kuprel B，Novoa R A，et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature，2017，542(7639)：115-118.

[6] Karpathy A，Toderici G，Shetty S，et al. Large-scale video classification with convolutional neural networks[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014：1725-1732.

[7] Radford A，Metz L，Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434，2015.

[8] Koller O，Ney H，Bowden R. Deep hand：How to train a CNN on 1 million hand images when your data is continuous and weakly labelled[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：3793-3802.

[9] Nguyen A，Yosinski J，Clune J. Deep neural networks are easily fooled：High confidence predictions for unrecognizable images[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015：427-436.

[10] Farfade S S，Saberian M J，Li L J. Multi-view face detection using deep convolutional neural networks[C]. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM，2015：643-650.

[11] Levi G，Hassner T. Age and gender classification using convolutional neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015：34-42.

[12] Mansanet J，Albiol A，Paredes R. Local deep neural networks for gender recognition[J]. Pattern Recognition Letters，2016，70：80-86.

[13] Nguyen A，Dosovitskiy A，Yosinski J，et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[C]. Advances in Neural Information Processing Systems. 2016：3387-3395.

[14] Graves A，Mohamed A，Hinton G. Speech recognition with deep recurrent neural networks[C]. Acoustics，speech and signal processing (icassp)，2013 IEEE international conference on. IEEE，2013：6645-6649.

[15] Grover A，Kapoor A，Horvitz E. A deep hybrid model for weather forecasting[C]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2015：379-386.

[16] Hossain M，Rekabdar B，Louis S J，et al. Forecasting the weather of Nevada：A deep learning

approach[C]. Neural Networks (IJCNN), 2015 International Joint Conference on. IEEE, 2015: 1-6.

[17] Torres C A, Crowther M R, Brodsky S. Addressing Acculturative Stress in Psychotherapy: A Case Study of a Latino Man Overcoming Cultural Conflicts and Stress Related to Language Use[J]. Clinical Case Studies, 2017, 16(3): 187-199.

[18] Lawrence J A. Positioning Children in Developmental Analyses of Cultural Conflicts[J]. Human Development, 2016, 59(2-3): 152-161.

[19] Chekired D A. Smart Grid Solution for Charging and Discharging Services Based on Cloud Computing Scheduling[J]. IEEE Transactions on Industrial Informatics, 2017.

[20] Jamali S, Alizadeh F, Sadeqi S. Task Scheduling in Cloud Computing Using Particle Swarm Optimization[J]. The Book of Extended Abstracts, 2016, 192.

[21] Guatieri S, Badaracco G, Defilippis I, et al. Pulse induction parking sensor[C]. SENSORS, 2016 IEEE. IEEE, 2016: 1-3.

[22] Kalantar B, Mansor S B, Sameen M I, et al. Drone-based land-cover mapping using a fuzzy unordered rule induction algorithm integrated into object-based image analysis[J]. International Journal of Remote Sensing, 2017, 38(8-10): 2535-2556.

[23] Narayanan M, Cherukuri A K. A study and analysis of recommendation systems for location-based social network (LBSN) with big data[J]. IIMB Management Review, 2016, 28(1): 25-30.

[24] Patil N D, Bhosale D S. Providing highly accurate service recommendation for semantic clustering over big data[J]. 2017.

[25] Carnelli P E, Yeh J, Sooriyabandara M, et al. ParkUs: A Novel Vehicle Parking Detection System [C]. AAAI. 2017: 4650-4656.

[26] Zadeh N R N, Cruz J C D. Smart urban parking detection system[C]. Control System, Computing and Engineering (ICCSCE), 2016 6th IEEE International Conference on. IEEE, 2016: 370-373.

[27] McPherson R A, Pincus M R. Henry's Clinical Diagnosis and Management by Laboratory Methods E-Book[M]. Elsevier Health Sciences, 2017.

[28] Patterson T F, Thompson Ⅲ G R, Denning D W, et al. Practice guidelines for the diagnosis and management of aspergillosis: 2016 update by the Infectious Diseases Society of America[J]. Clinical Infectious Diseases, 2016, 63(4): e1-e60.

结　束　语

深度学习模型之所以能够火起来,得益于大数据技术的发展,正是因为大数据技术,使得海量数据的深度学习成为可能。同时深度学习模型的发展,势必进一步促进大数据智能处理能力的提高。深度学习与大数据的共同作用是产生智能,为人类服务。

深度学习离不开大数据,因为如果没有海量的数据样本进行训练,深度学习模型无法获得很高的准确率,同时如果没有大数据的处理能力,深度学习模型的层数就会受到限制而无法达到一定深度,从而无法实现深度学习的功能。

大数据离不开深度学习,因为大数据与小数据不同的是大数据要挖掘的知识模式往往不明确,也就是说,难以用公式或者模型刻画,此时深度学习模型就有了英雄用武之地,因为深度学习模型中可以蕴涵极其复杂的难以刻画的模型。